**Content and collaborate based Sinhala Book Recommendation System**

P. N. C. Perera

University of Colombo School of Computing, Colombo 07, Sri Lanka

[nc21perera@gmail.com](mailto:nc21perera@gmail.com)

Prof. M.G.N.A.S. Fernando

University of Colombo School of Computing, Colombo 07, Sri Lanka

[nas@ucsc.cmb.ac.lk](mailto:nas@ucsc.cmb.ac.lk)

**Abstract**

The saying "Reading makes a man perfect" tells us how important it is to gain knowledge by reading different things like books, articles on the internet, newspapers, magazines, or even simple pieces of paper. This article talks about how reading a lot can have a big impact on people. It says that people who read regularly usually have a better understanding of life, make smarter choices, and handle difficult situations better. The traditional way of picking books, like choosing randomly or going by recommendations, can be hard, especially as libraries have more and more books. The COVID-19 pandemic and distance can make it even harder for people to find the books they want. This situation shows that there is a need for a better way for people to borrow books from the library. While many online platforms use systems that suggest products, these usually focus on making more sales and not necessarily on what the user really likes. This article talks about the lack of a system that suggests Sinhala books and suggests a new idea. The proposed solution combines both content and collaboration methods to create a hybrid model for recommending Sinhala books. This system aims to help online users find interesting and relevant books without wasting too much time and money. It not only benefits readers but also gives useful information to authors, helping them understand what readers like and adjust their writing to match popular interests. This recommendation system has the potential to change the way Sinhala books are discovered and chosen, making it a useful tool for both readers and authors in the digital age.

***Keywords:*** Sinhala Books, Collaborative filtering, Content based filtering, Hybrid model, Recommendation system

**1. Introduction**

With the increase of online shopping items, the need for giving the confidence in buying products has been increased. Many ecommerce companies like Amazon, Netflix, Spotify utilizing the power of recommendation system by implementing in their website to boost their sales. Even machine learning technique can be used to solve so many problems in real world, making product recommendation is widely known application of machine learning.

There are many recommendation systems have been implemented for many domains such as movies, books, e commerce items and songs. In this research the author is implementing a book recommendation system which help readers to find similar books or books read by other users who has similar preference as you. Since most applications have been implemented for English books and there is no system for Sinhala books, this research is focusing on implementing a book recommendation for Sinhala books.

People used to read from their childhood. As per the research, it says, People who read a lot tend to know more about life and are smarter when making decisions and handling difficult situations. (Marappan, 2022). In today’s world, time has more value and the researchers have no much time to spend on searching for the right articles according to their research domain. (Murali et al., 2019). Book readers usually select books by reading some random pages or asking someone to recommended any book. When reading that book, if he finds that the book is not interesting, he will not read any book after that. therefore, it is better to suggest books that he is interested in. With the increase in library collections, it is diﬃcult for readers to quickly ﬁnd the books they want. It is also diﬃcult for readers to ﬁnd Sinhala books of interest in a short period of time in the face of various bibliographies. Therefore, the user experience of the traditional library borrowing method is poor.(Dhanda and Verma, 2016) Due to the Covid-19 pandemic situation and the geographical barriers also it becomes a tremendous challenge for readers (Sarma et al., 2021) to find a relevant book as they do not like to go out and spend time searching books of their preference. Even the pandemic period is over it is better to be prepared to face such situation in future.

**2. Literature Review**

Several researches have been conducted related to Book recommendation systems and most of them have used machine learning techniques to build the model which is used to generate the list of items that user prefer.

In a study by (Sarma et al., 2021) researchers created a book recommendation system for online users. They used clustering to rate and suggest books based on similarity, utilizing data from Kaggle's Good readers book repository. The system identified and excluded potentially boring books through a classifier. Evaluation metrics included precision, sensitivity, specificity, and F1 score, along with a graphical accuracy representation through a receiver operating characteristic (ROC) curve. Future work aims to propose a system for recommending online courses using Convolutional Neural Network (CNN) technology.

The research (Wadikar et al., 2020) introduces a platform using a Convolutional Neural Network (CNN) for book recommendations through text processing and image classification. Text input from users is processed, and a dataset is collected by web scraping Amazon and Flipkart. In image classification, users upload book cover images for results. Cosine similarity is used to find related books. Advantages include no need for feature engineering, efficient handling of unstructured data, and quick access to highly-rated books. Evaluation and validation processes are lacking in the research.

The study by (Shah, 2019) created an e-commerce application employing collaborative filtering algorithms. Users could input ratings or sentences, and the system used natural language processing to calculate ratings based on sentence nature. The research delved into challenges like Scalability, Sparsity, Security, Cold start, and veracity in recommendation systems. Various methods were discussed, including clustering, classification, and item-based collaboration. The study used the "goodbooks10k" dataset on Kaggle, experimented with Python, and assessed accuracy using Mean Absolute Error (MEA).

Top of Form

**3. Theoretical Concepts**

***3.1 Content based***

A content-based book recommendation system suggests books to users based on the characteristics of books they have already liked. It works by analyzing the content or features of books, such as genres, authors, and keywords, to find similarities with the user's preferences. For example, if a user enjoys mystery novels with a specific author, the system will recommend other mystery books by the same author or within the same genre. This approach relies on understanding the intrinsic qualities of books and matching them to the user's known preferences, making it a personalized recommendation system that doesn't require information about other users.

***3.2 Collaborate based***

A collaborative-based book recommendation system works by analyzing the preferences and behaviors of a group of people to suggest books. Instead of relying on individual preferences alone, it considers the collective tastes and choices of a community. If people with similar reading habits enjoyed a particular book, the system might recommend that book to others with comparable tastes. This approach leverages the wisdom of the crowd, making recommendations based on the shared preferences of a community rather than focusing solely on an individual's likes and dislikes. It's like getting book suggestions from a group of friends who have similar reading interests, making the recommendations more likely to match your taste.

***3.3 Sentiment Analysis***

Sentiment analysis is like teaching computers to understand feelings in written text. Imagine you write something online, like a review or a tweet, and the computer reads it to figure out if you're happy, sad, angry, or neutral. It's like a digital detective for emotions! This helps businesses know how people feel about their products or services, and it also allows social media platforms to filter out negative or harmful content. So, sentiment analysis is a smart way for computers to catch the vibe of what people are saying online.

***3.4 Artificial Neural Network***

Artificial Neural Network (ANN) is kind of a virtual brain inspired by the way our own brains work. It's a computer system designed to learn and make decisions, just like humans do. In this network, there are nodes, or artificial neurons, that are connected in layers. These connections have weights that determine the strength of the relationship between the neurons. During training, the network learns from examples, adjusting these weights to improve its ability to make predictions or classifications. It's like teaching the network to recognize patterns and make sense of information. Once trained, the ANN can take in new data and use its learned knowledge to make predictions or decisions. So, in a nutshell, an Artificial Neural Network is a computer system that learns and thinks in a way inspired by the human brain to solve problems or make decisions.

**4. Methodology**

***4.1 Data Collection***

Since there are no data set available in the data set provides like Kaggle, a google form was shared among book readers groups in order to collect considerable amount of dataset for the research. 845 users contributed by filling the form with 5 books each and total of 4225 records could be collected. There are more than 300 popular books were considered for the research and each book details were collected via online book stores and saved in the database in order to display the book details once a book is selected by a user in the application.

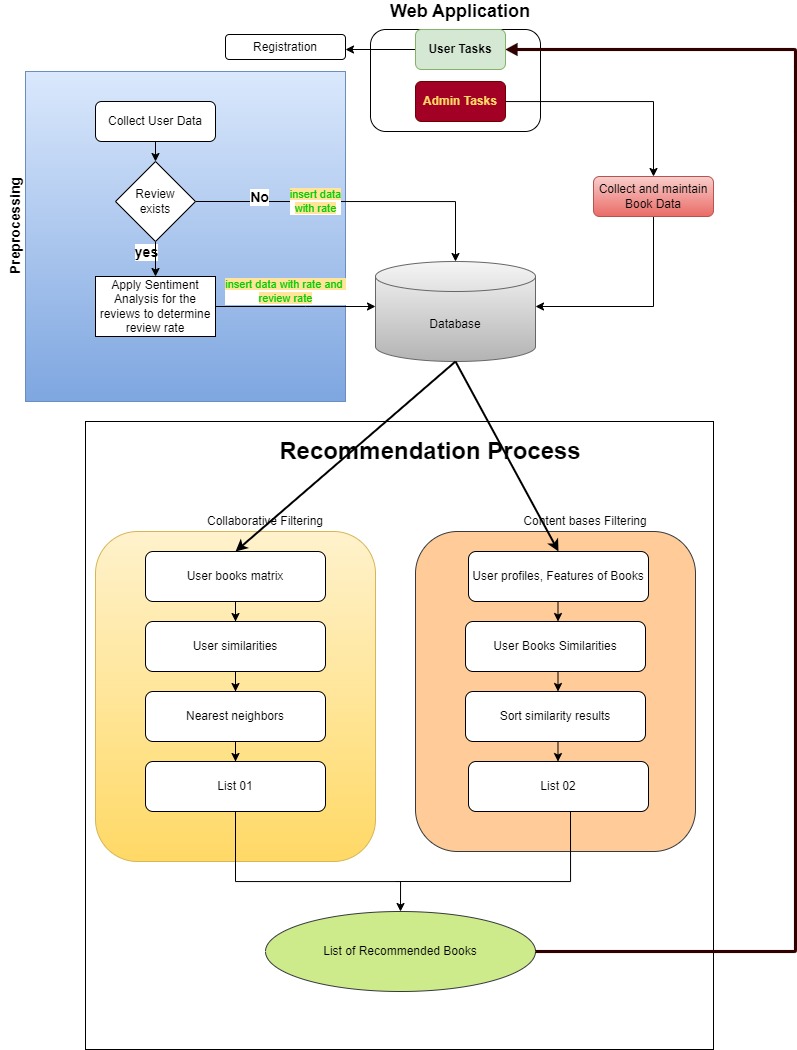
***4.2 Preprocessing***

In the google sheet shared to the user, there were two significant fields name rate and review. The rate could be selected in a range from one to ten and based on the provided rate a user can enter a review considering why you have given such rate for the selected book. Since a user can select a good rate and provide a negative review, a mechanism had to be used to overcome the problem. That is applying sentiment analysis to the review and generate a new rate based on the review provided. The final rate was considered by calculating the mean value or user given rate and review rate. Additionally, two separate models for content based and collaborate based have been created to predict the final book list.

***4.4 Technology stack***

The recommendation system was implemented using Python and popular libraries such as python Flask, Google translator in python, nltk and scikit-learn.Flask Python library with html, css and bootstrap ware used to implement the web application and scikit-learn, nltk, google translator in python were used in preprocessing and main business logics. Mysql workbench was used as main database to store and maintain the data

***4.3 Architectural Diagram***

.

In this system, data is first collected and then organized into three tables: Book\_Details, User\_Details, and Rate\_Details. User information is stored in User\_Details, while Book\_Details contains additional book details. Ratings given by users for books are stored in Rate\_Details, and if there's a review, a separate review rate is assigned after sentimental analysis and included in Rate\_Details. The final calculation involves taking the mean value of review and normal rates. Data preprocessing is essential before recommendation, removing null rate values and eliminating instances of maximum rates for a single book.

The recommendation process involves applying algorithms, with Collaborative Filtering being a primary method. This method recommends books based on similarities found from other users. User-based collaborative algorithms involve building a matrix of users and books with ratings, computing cosine similarity, and generating a list of recommended books. Content-based approaches focus on book descriptions and user preferences, creating a user preference vector based on past ratings. Hybrid methods, combining collaborative and content-based filtering, are commonly used for more effective book recommendations. The final recommended book list is generated by displaying the common books from collaborative and content-based filtering.

fig 1: Architect Diagram

The system implementation also addresses challenges, such as the increasing number of research papers. Efficient searching and filtering mechanisms, as mentioned in Murali et al. (2019), are crucial to save researchers' time. The paper highlights the use of algorithms to recommend quality research papers based on users' queries and preferences.

In conclusion, this system involves collecting, categorizing, and processing data to recommend books efficiently. It employs Collaborative Filtering, Content-Based Filtering, and a Hybrid approach for effective recommendations. The process includes data preprocessing, algorithm application, and a final recommendation list generation, enhancing the overall user experience in finding relevant books.

**5. Results and Discussion**

***5.1 Results of PCA and MLR Model***

Table 3. shows the results of the KMO and Bartlett’s sphericity test. Overall Kaiser’s measure of sampling adequacy is equal to 0.780, indicating that the sample size is adequate to apply the PCA. The significance value of Bartlett’s sphericity test is less than 0.05 and it also implied that PCA is applicable to our data set (P<0.05).

According to the results of PCA as shown in Table 4., out of 17 principal components only 4 principal components (PC1-PC4), explaining 84.498% of variance, were retained.

The principal component scores of selected PCs (PC1-PC4) shown in Table 5. are used as predictor variables for MLR. According to Table 5., Male & Female population, 0-19 &20 and above population, categories of education attainment and employment status are belong to PC1, Mean household income, Food and Non-Food Expenditure belong to PC2, Weather attributes (Rainfall, Temperature and Humidity) belong to PC3 and date belong to PC4.

|  |
| --- |
|  |
|  |

Table 7. Model Summary

Fig. 2, 3 and 4 shows the histograms indicating the distribution of residuals for the three dependent variables.

Fig. 2. Histogram of Frequency vs Regression Standardized Residual for Non-Biodegradable waste

Fig. 3. Histogram of Frequency vs Regression Standardized Residual for Biodegradable waste

Fig. 4. Histogram of Frequency vs Regression Standardized Residual for Total waste

***5.2 Results for the ANN Model***

For all modeling cases, all possible combinations of activation functions between the hidden layer and output layer have been tested and the results are tabulated in Table 8. According to Table 8, the hyperbolic tangent function for the hidden layer and sigmoid function for the output layer shows the least errors.

Table 8. SSE for different combinations of activation functions

*5.2.1 Optimum number of neurons in the Hidden Layer*

After choosing the activation functions, the optimum number of neurons in the hidden layer is found by running the ANN by varying the number of hidden neurons from one to nineteen. Table 9 presents the SSE values for training and test sets and average overall relative error values for training, test and validation sets,while varying the number of hidden neurons.

Fig. 5. Sum of squared Training error vs. No. of neurons in the hidden layer

Fig. 6. Average Overall Relative Error of Training, Test and Validation sets vs. No. of neurons in the hidden layer

Figure 5 and 6 indicates a graphical representation of the same. According to the similarity of the above figures, it is proved that when there are 8 hidden neurons, SSEs and Average Overall Relative Error for training, test and validation sets are minimized. Further, test SSE, starts to go up after 8 neurons, possibly indicating overfitting. Therefore, optimum number of hidden neurons for the ANN structure is 8.

*5.2.2 Features of the selected ANN Structure*

Table 10 presents the network information of the chosen network structure. In the architectural point of view, it is a 19-8-3 neural network (19 independent variables, 8 neurons in the hidden layer and 3 dependent variables).

Table 10. Network Information

Fig. 7. shows a complete connected graph of input, hidden layer and output respectively. The structure in the Fig. 7. is a feed forward architecture because the connections in the network flow forward from the input layer to the output layer without any feedback.

Fig. 7. Multi-Layer Perceptron Architecture Structure

*5.2.3 Summary of the ANN Model*

The following model summary in Table 11 shows information about the results of the neural network training. Here, SSE is shows because the hidden and output layers use the hyperbolic tangent and sigmoid activation functions, respectively. This is the error function that the network tries to minimize during training.

Table 11. Model Summary

|  |  |  |  |
| --- | --- | --- | --- |
| Training | Sum of Squares Error | | 14.369 |
| Average Overall Relative Error | | .136 |
| Relative Error for Scale Dependents | Nonbiodegradable | .160 |
| Biodegradable | .149 |
| Total | .102 |
| Stopping Rule Used | 1 consecutive step(s) with no decrease in error | |
| Training Time | | 0:00:01.97 |
| Testing | Sum of Squares Error | | 4.300 |
| Average Overall Relative Error | | .119 |
| Relative Error for Scale Dependents | Nonbiodegradable | .138 |
| Biodegradable | .135 |
| Total | .087 |
| Holdout | Average Overall Relative Error | | .125 |
| Relative Error for Scale Dependents | Nonbiodegradable | .146 |
| Biodegradable | .134 |
| Total | .097 |

The scatter plot of the observed against predicted for the test data set are given in Fig. 8, 9 and 10 for the Non-Biodegradable, Biodegradable and Total waste respectively, for the ANN model. R2 values obtained are 0.846, 0.855 and 0.902 for Non-Biodegradable, Biodegradable and Total waste, respectively.

Fig. 8. Graph of predicted vs observed for Non-Biodegradable waste

Fig. 9. Graph of predicted vs observed for Biodegradable waste

Fig. 10. Graph of predicted vs observed for Total waste

*5.2.4 Importance of Independent Variables*

Table 12 shows the results of the Variable Important Analysis, which computes the importance and the normalized importance of each variable in determining the neural network. Fig. 11 shows a graphical representation of the normalized importance of predictor variables. The analysis is based on the training and testing samples. The importance of an independent variable is a measure of how much the network’s model-predicted value changes for different values of the independent variable. Moreover, the normalized importance is simply the importance values divided by the largest importance values and expressed as percentages. From the following table, it is evident that GCE(AL) population contributes most in the neural network model construction, followed by economically not active and Degree and above population.

Table 12. Independent Variable Importance

|  |  |  |
| --- | --- | --- |
| Variables | Importance | Normalized Importance |
| Male | .059 | 49.7% |
| Female | .044 | 37.3% |
| @019 | .063 | 53.6% |
| @20ampabove | .041 | 34.2% |
| Primary | .021 | 17.6% |
| Secondary | .024 | 20.2% |
| GCE(OL) | .070 | 59.0% |
| GCE(AL) | .119 | 100.0% |
| Degreeandabove | .113 | 95.5% |
| Noschooling | .028 | 23.4% |
| Unemployed | .031 | 26.3% |
| Employed | .083 | 70.4% |
| Economicallynotactive | .114 | 96.5% |
| Meanhouseholdincome | .026 | 22.1% |
| Foodexpenditure | .064 | 54.3% |
| Nonfoodexpenditure | .023 | 19.5% |
| obs\_valmm | .006 | 4.8% |
| Tempmax | .040 | 33.9% |
| RelativeHumidity | .030 | 25.4% |

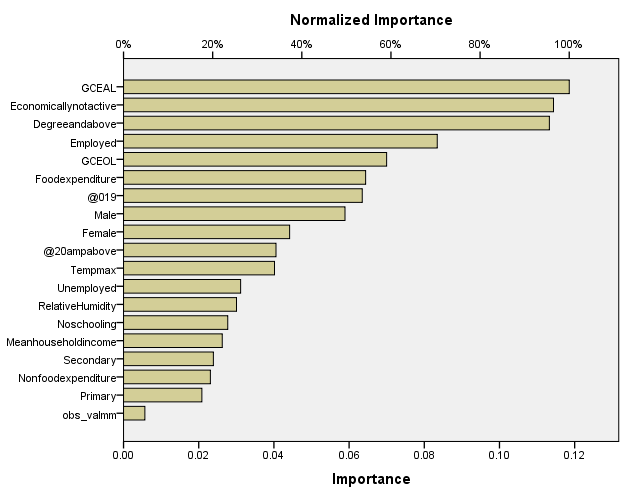


Fig. 11. Normalized Importance

**6. Conclusion**

This study presents a systematic process to identify the significance of factors affecting waste generation and a methodology for developing solid waste generation models with various socioeconomic, demographic and geographic variables and climatic factors.

PCA was carried out to investigate the influencing factors to waste generation and to avoid the effects of multicollinearity among them. Additionally, the regression relationships for estimating waste generation, based on the selected key factors from the PCA, are developed. The PCA shows four components of key factors that can explain at least up to 84.498% of the variation of all variables. Then a MLR analysis carried out with the factor scores obtained from PCA which showed R2 values of 0.750, 0.544 and 0.769 for Biodegradable, Non-Biodegradable and Total waste, respectively. Neural Network model is best fitted with R2 values of 0.846, 0.855 and 0.902 and lower RE values of 0.138, 0.135 and 0.087 for Non-Biodegradable, Biodegradable and Total waste, respectively. Therefore, ANN model which showed higher predictive accuracy, is concluded as the appropriate model. Further, it is concluded that GCE (AL) population contributes most in the ANN model construction, followed by economically not active and Degree and above population.

Data availability is a limitation of this study, however, despite the limited data, the proposed model reached satisfactory R2 and lower error values and learnt to model the desired output with a good accuracy. This study provides a reliable method for estimating solid waste generation, providing decision makers, useful information for waste management policy development.

**7. References**