A Natural Language Processing Approach to a Skincare Recommendation Engine

MSc Research Project
Programme Name

Adelola Adebo
Student ID: X19134711

School of Computing
National College of Ireland

Supervisor: Dr Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Adelola Adebo

Student ID: X19134711

Programme: MSc Data Analytics **Year:** 2020

Module: MSc Research Project

Supervisor: Dr. Vladimir Milosavljevic

Submission Due Date: 17th August 2020

Project Title: A Natural Language Processing Approach to a Skincare Recommendation Engine

Word Count: 8014 **Page Count** 25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……………………………………………………………………………………………

Date: ……………………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Natural Language Processing Approach to a Skincare Recommendation Engine

Adelola Adebo
X19134711

Abstract

Recommendation systems play a significant role in helping users to narrow their choice. Traditional recommendation techniques require the user's rating history to predict unknown ratings. Recently, a new line of recommendation system research has emerged, which seeks to exploit user reviews to serve as an alternative source of recommendation knowledge. In this study, we use the information included in the user reviews to develop a skincare recommendation system. The proposed methodology seeks to combine different review elements to address the issue of rating prediction. Another advantage of our proposed system is that it uses a deep neural network to recognize the commonalities between users and items. It also uses a word embedding and an additional Long-Short-Term Memory encoder to learn better semantic word information instead of using traditional models such as bag-of-word models. The efficiency of the proposed framework in rating predictions was assessed using the Mean Squared Error Metric. Experimental findings show that the new recommendation method has performed better than the baseline approaches and can be used for e-commerce purposes.

1 Introduction

In recent years, there has been a significant increase in the variety and number of skincare products provided by companies. Having too many options makes it overwhelming for customers to identify which product is right for them when they shop for skincare products in-store or online. With the rise of online social sites, people now have access to information and product reviews freely shared by similar users to help in their decision making. Forum websites like Reddit, and online retail websites such as Amazon have a massive amount of user reviews where people express their opinions about products and services. Therefore, in this study, we aim to leverage the rich and valuable knowledge contained in online reviews to develop personalised skincare recommendations. Personalised recommendations can help customers find the right product, save time and improve their shopping experience while at the same time assisting the brands to increase their sales and conversion rates.

Many of the widely applied recommendation techniques are based on collaborative filtering (CF) approach. CF has been used for a variety of applications, e.g. in recommending movies on Netflix, songs in Spotify and products on Amazon. CF makes recommendations to users based on the preferences of other users with similar tastes. The matrix factorisation (MF) technique has proven to be the most effective of collaborative filtering approaches. Past ratings are critical for this technique to learn the commonality between items and users. Despite their success, their widespread use has revealed some shortcomings, one of these shortcomings is seen in cases where rating data are sparse or unavailable, which may lead to poor recommendations. Another weakness is that CF techniques do not capture the underlying reason for the user’s rating, and as a result, cannot accurately capture the taste of a user.

A new direction in the recommender system's research has suggested using user reviews to overcome the limitations associated with CF because it would reduce the need for explicit ratings. Specifically, some researchers have argued that using textual user reviews can substantially improve prediction accuracy in cold-start settings where the system has not gathered sufficient information about the user or item to generate recommendations. This is because user reviews are believed to be much more expressive and explain the user's preferences towards a product compared to numerical ratings, hence, they can contain additional information that can be used when there are insufficient user ratings (Chen et al., 2015).

As a consequence of the merits of text reviews, many researchers have used it for recommendations. Chen et al. (2015) defined different review elements that can be derived from user reviews and used by recommender systems, among these review elements, review words, i.e. terms, opinions, and topics have shown to be useful for review analysis. Some researchers conducted sentiment analysis on user reviews to derive the user's overall opinion (positive, negative, or neutral) or an opinion on the product aspect (Zhang et al., 2013), the extracted opinion can be transformed into numerical ratings to improve CF techniques. Other researchers have used topics. Topics relate to aspects of the item that a user discusses in a review, e.g. "the sound quality of the phone is great." Phone and sound quality are the topics under discussion. Different topic modelling techniques such as Latent Dirichlet Allocation (LDA) can be used to extract these topics, after obtaining the topics, implicit scores inferred from the topics can be integrated into the standard CF techniques to improve recommendations (McAuley and Leskovec, 2013). One drawback of topic-based approaches is their failure to capture word context information which has proven invaluable to successful recommendation systems.

The rise of sophisticated deep neural network models has made it possible to better capture and integrate review words into existing CF models. Zheng et al. (2017) used Convolutional Neural Networks (CNN) to learn the latent representations of users and items from their review texts, the learned hidden factors are then combined into a rating predictive MF machine. While deep learning techniques have shown to achieve state-of-the-art recommendation algorithms, some of the approaches fail to fully consider the user opinion when modelling the user and item interaction. Besides, some use only one form of review element.

This research uses a novel approach to personalised skincare recommendation systems. The proposed model uses different review elements to jointly model user and item interaction to address the problem of rating prediction. The proposed model learns common latent factors for users and items using three neural networks aligned to maximise rating predictions. One network learns the characteristics of the user, the other learns the properties of the item, and the third - a Long Short-Term Memory (LSTM) model - learns the contextual information of words, the learned hidden features for users and items are then used to predict ratings.

The proposed approach will help us in addressing the following research question:
RQ: How effectively can the text analysis of online user reviews be used to develop skincare recommendations?

In addressing the research question, the objectives of this research are listed below:

1. Investigate skincare recommendation techniques.
2. Critically review state-of-the-art review-based recommendation systems.
3. Conduct experiments to determine the effect of different review elements.

The rest of the paper is organised as follows: Section 2 discusses the related works that have been done in this area of research, section 3 provides a description of our proposed approach, while section 4 sets out the implementation of the study. Section 5 describes the experiments done, analyses of the experiments and then the findings. Section 5 also illustrates how the research's aim and objectives were achieved, it details how it differs from and improves on previous related works. The conclusion of this research work is provided in Section 6.

2 Related Work

This section discusses related approaches to this research and how they differ from our approach, it is divided into two categories including:

1. Existing work on skincare recommendation systems and
2. Existing work on review incorporated recommendation systems.

It focuses on the motivation, the theoretical essence of each technique, the findings of the study and the comparison of the different algorithms used in each recommendation method.

2.1 Skincare recommendation systems

A few research works have attempted to develop a skincare recommendation system, of these, Nakajima et al. (2019) tried to develop a personalised skincare recommendation system by incorporating user reviews into a content-based recommendation algorithm. First, the author's built user profiles based on their age, skin type and desired beauty effect (e.g. moisturising, anti-ageing, organic cosmetics) obtained from the website data used, cosmetic products were then extracted from user reviews that fit each attribute of the user. Using the Ingredient Frequency-Inverse Product Frequency IF-IPF method, the system ranked the desired ingredient and recommended products containing these ingredients as their main ingredients. Their experimental results showed that their model was generally good.

Recently, Lee (2020) adopted a similar approach to Nakajima's research, he used the IF-IPF ingredient ranking method to filter products and to recommend appropriate products based on the user's skin type. Nakajima et al. (2019) and Lee's (2020) both investigated product recommendations based on the composition of ingredients in the products, however, Lee's work did not incorporate user reviews. One drawback to Nakajima et al. and Lee's research is the over-specialisation of the content-based recommendation approach. The content-based approach relies heavily on past user interest and, as such, limits the algorithm's ability to recommend new items. For example, if a cosmetic product is rated by the user, the algorithm would only recommend products containing similar ingredients. With this approach, the user would never discover anything new. In this research, the proposed method leverages a collaborative filtering approach and identifies a group of users with similar tastes. Users can thus benefit from a variety of products recommended by related users.

Another drawback to Nakajima et al. (2019) and Lee's (2020) works is that they did not use a machine learning approach to recommend products, hence, the system couldn't learn how to recommend products by itself, besides, when modelling the user profile, they ignored the users sentiments toward the ingredient. To the best of our knowledge, our paper would be the first to develop a personalised skincare recommendation system using machine learning techniques.

2.2 Review-based recommender system

This subsection discusses existing research on review-based recommender systems. Chen et al. (2015) defined different review elements that can be derived from user reviews and used by recommender systems. Among these review elements, review words, i.e. terms, opinions, and topics have shown to be useful for review analysis. For clarity, we divide this subsection into two categories. Firstly, we discuss techniques that build a user profile, product and ratings profile by mining user reviews and secondly, we discuss deep learning approaches to review incorporated recommendation systems.

.

2.2.1 Techniques based on review terms and inferred ratings

One of the earliest studies, Garcia Esparza et al. (2012), extracted frequently used terms from user reviews to build user profiles and product profiles, the extracted keywords were used to create an index representing users and products. The target user's index during the recommendation process served as a query that matched the product index and was used to produce a list of the most related products. The proposed approach was evaluated with the traditional rating CF-based method, while its accuracy was marginally less than that of the rating-based CF approach, its findings indicated that it was superior in terms of the diversity of products recommended.

Sentiment analysis is another approach used by researchers to augment the rating prediction problem. The main goal is to determine whether a review's disposition is positive or negative or neutral (Srifi et al., 2020), for instance, in a study conducted by Ramzan et al. (2019), a user's sentiment towards an item was inferred by using his overall opinion which was derived by aggregating the sentiment score of all the opinion words in the user review. The inferred ratings were then used in place of real ratings for predictions. Similarly, Zhang et al. (2013) performed sentiment analysis to derive ratings and fused it into a matrix to generate recommendations, they compared their proposed method with virtual ratings and another model with real ratings, they reported that the inferred rating model performed better than the actual rating model.

The work of Qui et al. (2016) also shows the use of both sentiment analysis and keyword extraction to construct user and products profiles. Through the use of sentiment scores and item word frequencies, a user and an item attribute matrix are created. The learned user and item attribute matrix are then integrated with a rating matrix via a latent factor model and fed into a matrix factorisation machine.

Experimental results from the above approaches show that building users and product profiles from reviews can lead to improvements over traditional profiles built with static product details (Garcia Esparza et al., 2012). Also, the results from Zhang et al. (2013) show that the accuracy of CF techniques based on inferred ratings is similar to real ratings. This can be very useful in practice if there are no explicit ratings available.

In summary, these studies validate the potential of exploiting user reviews to capture more fine-grained user preferences and to enable CF to solve the problem of rating sparsity.

2.2.2 Techniques based on deep learning and matrix factorisation

Research work in this area focuses mostly on the methodology used to model user reviews and incorporate them into the matrix factorisation equation. In Almahairi et al. (2015), a probabilistic model was proposed based on paragraph vectors (PV-DBOW) and matrix factorisation. It learned the word representations of a review and incorporated the word vectors with ratings as a regularisation term. This approach is shown to improve over a baseline model that uses ratings on its own. A similar procedure was followed in Alexandridis et al. (2019). The main difference is that the authors used Paragraph vectors to learn word representations based on the context of the review so that the model could identify the similar words and phrases with a different meaning. However, the bag-of-words models used in the works described above cannot accurately model the semantic information of a text review, leading to a poor understanding of the semantic meaning of the review.

Other researchers applied autoencoders to model user reviews. In Wang et al. (2015), a stacked denoising autoencoder is used to extract deep feature representations from content information and to learn the implicit interactions between users and items. This network was combined with a probabilistic matrix factorisation model to calculate rating prediction for a target user. Another example is presented in Shoja and Tabrizi (2019), the authors employed LDA text analysis on user reviews to create a user-attribute matrix. The user-attribute matrix resulted in a very sparse matrix, to overcome this limitation, the authors used an autoencoder to learn latent features from the sparse matrix and reconstructed it to a dense matrix. The dense user-attribute matrix was then used alongside a user-item matrix for recommendation using Matrix factorisation. However, autoencoders represent words in a low-dimensional bag-of-words vector, they do not take the context of the word and word order into consideration (Zhang et al., 2019).

Zhang et al. (2016) tried to preserve word order by modelling user reviews using a multi-level embedding framework. The first layer used pre-trained word embeddings to capture word context information while the second layer modeled users and item embeddings. In the last layer, the model concatenated the user, item and review embedding layers and fed it into a matrix factorisation model for the rating prediction.

Another work that attempted to better capture word semantics from user reviews is in (Da'u and Salim., 2019). The authors proposed an LSTM encoder to learn user sentiments from text reviews. Long-Short Term Memory is a type of Recurrent Neural Network (RNN) that is widely used to learn patterns in sequences of data, such as text.

Kim et al. (2016) proposed a Convolutional Matrix Factorisation Model (ConvMF) to learn word semantics and contextual information better. The model uses CNN to perform convolution and max-pooling to learn the semantics of the review. After that, the learned latent vectors are combined into a probabilistic matrix factorisation model for rating prediction. The model proposed is shown to be superior to bag-of-words models.

Deep Cooperative Neural Network (DeepCoNN) (Zheng et al., 2017), a recent state-of-the-art model used CNNs to exploit user reviews. Two parallel CNN models were used to model users and items together. The first CNN learnt users' vector representations, while the second CNN learnt the items' vector representations. The learned user and item vector representations were coupled via a shared regression layer and fed into a factorisation machine for rating prediction. A limitation to DeepCoNN is that it requires a user's review for every item which is unrealistic in practice because products are supposed to be recommended to users before they experience it. To address this limitation, Catherine and Cohen (2017) proposed TransNets, an extension of DeepCoNN. TransNet introduced an additional latent layer that predicts the rating of the current review, which may be substituted if the original review were unavailable.

It is believed, however, that CNN has a word ambiguity problem due to the often too small sliding windows. Topic modelling was integrated with Recurrent Neural Networks (RNN) in Jin et al. (2018) to improve the retention of contextual information. RNN may retain context information, while the topic modelling may provide a word of co-occurrence to supplement in the case of loss of information. Finally, the two latent vectors were integrated into the matrix factorisation algorithm.

Despite the progress of deep learning techniques, the majority of the above mentioned works only used them to model auxiliary information. The primary interaction between users and items is still modelled by a matrix factorisation machine which may not be able to capture the complex non-linearities of the interaction between users and items. Also, matrix factorisation makes it challenging to integrate side features beyond the user's ratings (Zhang et al., 2019). NCF (He et al. 2017), proposed an end-to-end deep learning architecture to model the interaction between users and items using a Multilayer Perceptron. Their experimental results showed that deep models outperform vanilla matrix factorisation models. DNNRec of (Kiran et al. 2019), used embeddings to learn the non-linearities between users and items and incorporated side information. These embeddings were fed into a deep neural network for product recommendation. The author also investigated how learning rates can be optimised to improve prediction. In Zhang et al. (2017), deep learning techniques such as doc2vec and CNN were used to jointly learn embedding representations from text, image and ratings to improve recommendations.

2.2.3 Conclusion

Upon review of the recent works in review-based CF recommender systems, we note that further work is required. For example, fusing various elements of user reviews may be more effective in discovering more nuanced patterns in user preferences to improve recommendations. In addition, a personalised skincare recommendation system has not yet been developed, therefore, in this research, we explore user reviews beyond existing literature to develop a personalised skincare recommendation system. In our proposed framework, firstly, product keywords are extracted from user reviews to represent products, Secondly, we perform a sentiment analysis of the reviews to determine the user's ratings for the product, thirdly, given user reviews, product, and inferred ratings, a user-item matrix is constructed. Finally, a deep neural network is used to model user-item interaction to solve the problem of rating prediction. Table 1 shows a brief summary of the related works.

Table 1: Summary of literature review

Literature	Review elements	Recommending Method
Garcia Esparza et al. (2012)	Frequent terms	Content-based filtering
Ramzan et al. (2019)	Aspect sentiment scores and explicit ratings	Matrix factorisation
Almahairi et al. (2015)	Review words + explicit ratings	Word embeddings + matrix factorisation
Zhang et al. (2013)	Sentiment scores as inferred ratings	Matrix factorisation
Qui et al. (2013)	Frequent terms + Sentiment scores + explicit ratings	Matrix factorisation
Nakajima et al. (2019)	Frequent terms	Cosine similarity
Alexandridis et al. (2019)	Review words + explicit ratings	Words embeddings + matrix factorisation.
Wang et al. (2015)	Review words + explicit ratings	Autoencoders + matrix factorisation
Shoja and Tabrizi (2019)	Review words + topics + explicit ratings	LDA + Autoencoders + matrix factorisation
Zhang et al. (2016)	Review words + explicit ratings	Word embeddings + matrix factorisation
Da'U and Salim (2019)	Review words + topics + explicit ratings	Supervised LDA + LSTM + Attention + matrix factorisation
Kim et al. (2016)	Review words + explicit ratings	CNN + probabilistic matrix factorisation.
Zheng et al. (2017)	Review words + explicit ratings	2 CNN + factorisation machine
Jin et al. (2018)	Review words + topics + explicit ratings	LDA + LSTM + matrix factorisation
He et al. (2017)	Explicit ratings	Embeddings + Neural networks
Kiran et al. (2020)	Review words + explicit ratings	Embeddings + Neural networks
Our approach	**Review words + frequent terms + sentiment scores**	**Embeddings + LSTM + Neural network**

3 Skincare Recommendation Engine Methodology

This section explains the approach used to develop the Skincare Recommendation Engine. The system learns to model users and item interactions for collaborative filtering using text reviews. The research methodology follows the Knowledge Discovery in Databases (KDD)[1] approach for data mining projects as shown in Fig. 1. In order to develop the model, the following assumptions were made:

1. User-written reviews may be used to represent items and users.
2. The user reviews may imply the user's preferences for the item under discussion.

As a result of this, we formulated a three-step approach:

1. Identify the topics mentioned in the reviews.
2. Perform sentiment analysis to derive user ratings inferred from sentiment scores.
3. Model user and item interaction using a deep neural network.

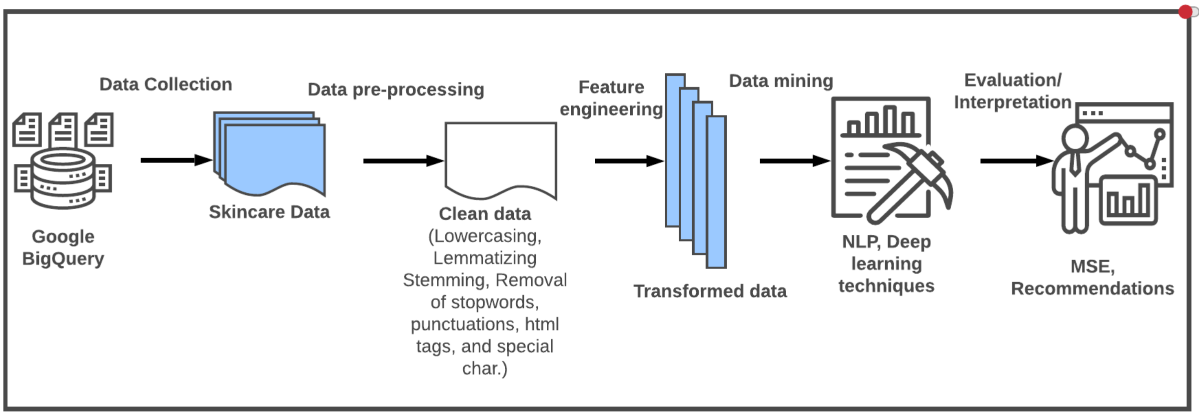

Figure 1: Proposed approach in line with KDD framework

3.1 Data Collection

In this research, we selected three datasets to develop the skincare recommendation system. The primary dataset is user comments from Reddit.com - a social media network. The second dataset is a skincare ingredient dictionary scraped from a trusted database, and the third is a skincare product dataset scraped from Boots.com[2] - a health and beauty retailer. All three datasets were ethically sourced.

Reddit is one of the most popular social networks where people share ideas, discuss issues of similar interest and ask questions, it has specific communities dedicated to particular themes called subreddits, specifically, the user comments in the r/SkincareAddiction[3] subreddit were used for this study. In the r/SkincareAddiction Subreddit, members typically discuss their skincare concerns, ask for or make recommendations, and share their review on various skincare products. The Reddit dataset is freely available on the Google Big Query published by Jason Baumgartner[4], a Reddit user. Between 2018 and 2019, 18 months of user comments were extracted and converted to csv format. It consists of over 1 million user reviews and aliases.

---

1 http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
2 https://www.boots.ie
3 https://www.reddit.com/r/SkincareAddiction/
4 https://www.reddit.com/user/Stuck_In_the_Matrix

The second dataset is a list of all cosmetic skincare ingredients. It was scraped from Paula's Choice repository[5]. It consists of 1,833 ingredients used in the formulation of skincare products and would be used to develop a skincare dictionary.

The third dataset was scraped from Boots.com and contains a list of skincare products. It contains 2,803 skincare products; it would also be used to match the products discussed in the comments of the Reddit user.

3.2 Data Pre-processing

Transforming text in natural language format to numeric representations is a critical component of the proposed methodology. A number of pre-processing tasks must be performed before building the model. At this stage the pre-processing done included removing HTML tags, special characters and punctuations, transforming all characters into lower case, splitting words into tokens, and finally, stemming and lemmatization to convert words to their root form.

3.3 Feature Engineering

Here, we extracted review elements and converted it to a user-item matrix that was used for modelling. The process flow for transforming text user reviews into a user-item matrix and using this matrix to predict ratings is shown in a schematic format in Fig. 1.

User/Product indexing: In this work, the concept of how to use text user reviews was by investigating what products/ingredients were mentioned in the user review. Drawing inspiration from Garcia Esparza et al. (2012), we extracted product keywords from text user reviews to represent products. To achieve this, we matched product keywords to a set of predefined products/ingredients.

Sentiment Analysis: Afterwards, we did sentiment analysis to derive scores that would serve as user ratings (following in the footsteps of Zheng et al., 2013). Since our data was unlabelled, we used an unsupervised approach to extract sentiments from the data. For this study, we used two sentiment lexicons, VADER (Valence Aware Dictionary and Sentiment Reasoner), a rule-based sentiment analysis model that is particularly suited to work well on social media data (Hutto and Gilbert 2014). VADER uses a list of lexical words that are scored according to their semantic orientation, ranging from the most negative sentiment-1 to the most positive +1. The second sentiment lexicon is SocialSent, a domain-specific sentiment lexicon for Reddit data (Hamilton et al., 2016). The VADER sentiment lexicon was updated with the domain-specific skincare Reddit lexicon obtained from SocialSent. Final sentiment scores were derived by summing up the sentiment scores for each word in the text.

[5] https://www.paulaschoice.com/ingredient-dictionary?crefn1=name-first-letter&crefv1=R

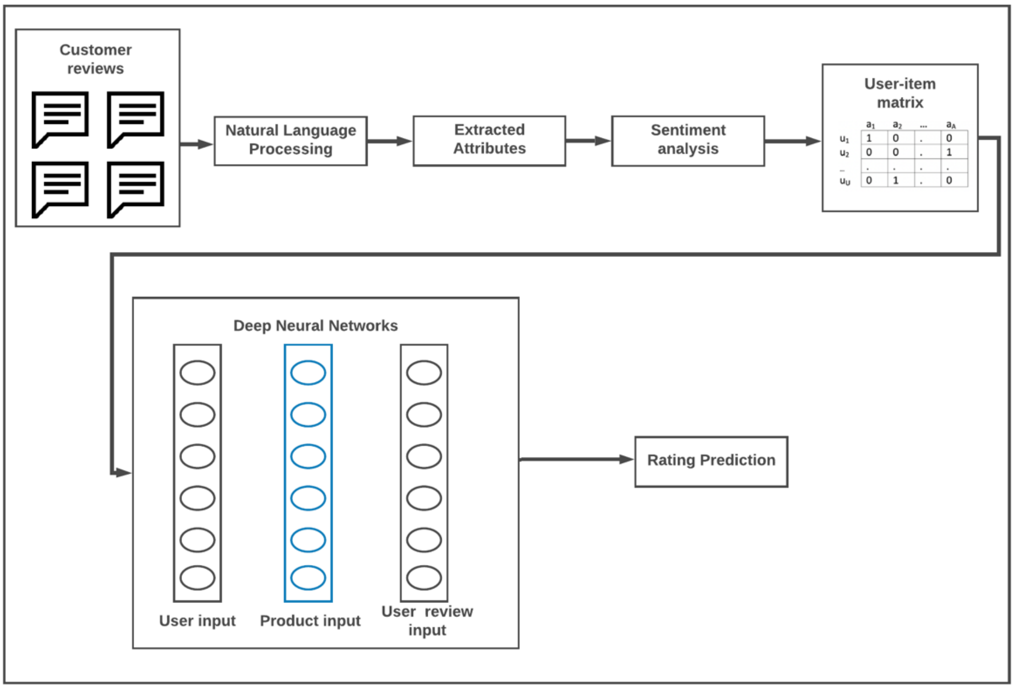

Figure 2: General structure of the proposed methodology in using user reviews for recommendations.

3.4 Recommendation Task

D is a corpus with a set of reviews. For a given item j written by a class of users u, and each review$D_{u,j} \in D$ is joined by an overall rating $r_{u,j}$ which depicts the user's feedback toward the items. Each user-item interaction can be denoted as a row $(u, j, r_{u,j}\ d_{u,j})$. The main aim is to calculate the unknown ratings$r_{u,j}$ of an item j on a user u that has not been disclosed by the user. Table 2 shows the notations used in this research.

Table 2: Notations

Notation	Definition
D	Corpus with a set of reviews
$r_{u,j}$	User u ratings for item j
$d_{u,j}$	User u review for item j
b_u	User bias
b_j	Product bias
W_u	User feature vector
V_j	Item feature vector
μ	Global average

3.5 Modelling (NCF + reviews)

To solve the problem in section 3.4, the proposed architecture in this research consists of four main parts including:

1. User Input
2. Product Input

3. The LSTM Layer – To model user text reviews and learn word semantics and contextual information better.
4. Predictive Rating Layer - For estimating the predicted ratings.

The following subsections provide a detailed description of the model.

3.5.1 User/Product Embedding Layer

Vector representations for the users and products were learned using an embedding layer. This is a better approach for transforming words into vectors than one-hot encoding as one-hot encoding leads to a sparse matrix that is difficult for machine learning.

3.5.2 Review Layer

The third input to the model is user reviews. Text reviews contain rich information about the user's preference towards the item. The text reviews were first converted to numeric sequences, and then vector representations were learned from the numeric sequences using word embeddings. Word embeddings take a given word from a document D and maps it to vectors of real numbers in a multi-dimensional space. In that space, each vector represents a given word. Typically, the embedding captures some of the semantic information of the input and places words with similar meanings besides each other. Subsequently, the output word vector sequences from the embeddings are passed into a Long Short-Term Memory (LSTM) layer. LSTM was adopted in this research because of its ability to learn contextual information across the entire input word sequence. The output of the LSTM layer is a latent document vector which is concatenated with the output vector from the user-item interaction.

3.5.3 Prediction Layer

The predictive rating layer is where the actual predictive rating function occurs for the recommendation process. The prediction layer accepts the user/item representations as input and the latent vector of the document D and uses them to predict ratings. The ratings can be derived as:

$$r_{u,j} = W_u V_j + b_u + b_j + \mu \quad \textbf{(1)}$$

where the user, item and global bias are added as regularisers to prevent outfitting.

3.6 Evaluation

The accuracy of the prediction of the recommendation system is the most widely discussed property in recommender system's literature. A prediction engine is used for most recommendation engines (Gunawardana, and Shani, 2015). This algorithm will forecast user preferences about products, in our case, skincare products. One underlying premise in a recommendation system is that the consumer wants a system that makes more precise predictions. Thus, most researchers have also sought to find algorithms with better forecasts. The mean squared error metric (MSE), one of the most popular metric, is adopted to assess the accuracy of the rating prediction model. Further experiments are also conducted to evaluate the effect of each of the component of the proposed system. A detailed explanation of the experiments and evaluation is given in Section 5.

4 Implementation

In this section, a description of the test environment and the implementation of the proposed model is done.

4.1 Technology Setup

Below is a list of the items we used for our technology setup:

1. Google Collaboratory from Google Cloud Platform GCP - This was used to implement the deep learning models used in this research. It was used because of the large dataset we had to pre-process as iterating through several matrices requires large amounts of computing resources.
2. MacBook Pro with the following specification:
    a. Core i5 Processor,
    b. 8 GB of RAM and
    c. 500 GB of HDD Hard Drive.
3. Python programming Language
4. Tensorflow and Keras - These are the deep learning libraries used.
5. Jupyter Notebook - This is part of the Anaconda framework; it was used for the overall implementation of the project both on the Google Cloud and the local machine used.

4.2 Data Cleaning and Pre-processing

The data cleaning and pre-processing steps implemented for this research are as follows:

1. The Reddit comments dataset used was extracted from Google Big Query, and the skincare ingredient and product datasets were scraped from websites using the BeautifulSoup web scraping library.
2. After obtaining the dataset, we use modules from the NLTK (Natural Language Toolkit) library to perform the pre-processing tasks described in Section 3. The unstructured nature of Reddit social media data required a lot of processing to be done.
3. A dictionary was created containing skincare ingredients and products.
4. We created custom functions parsing the reviews to identify products/ingredient from the dictionaries that have been created.
5. Following this, sentiment analysis was carried out on the reviews to infer ratings using the VADER sentiment analysis tool available in python.
6. Lastly, a user-item matrix was prepared from the resulting matches.

4.3 Exploratory Data Analysis

In order to have a better understanding of the pre-processed data before modelling, an exploratory analysis was carried out. Figures 3 and 4 below show the analysis.

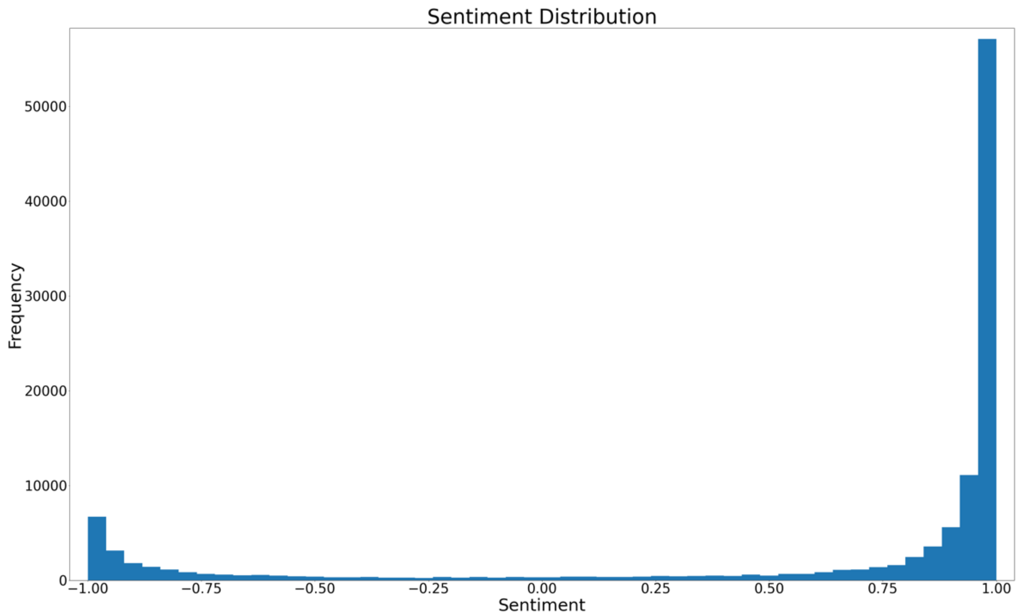

Figure 3: Sentiment Distribution

Fig. 3 shows the sentiment score distribution derived from the sentiment analysis performed on the user reviews. From this figure, we can see that the majority of ratings are overwhelmingly positive. From fig. 4, we were able to identify that the most popular skincare brand is "The Ordinary" and the most popular skincare ingredient is Hyaluronic acid. Other popular skincare ingredients mentioned include benzoyl peroxide, glycolic acid, salicylic acid, and micellar water.

Figure 4: Top Common words

4.4 Modelling

A total of four models with different inputs and layer types were designed with this architecture to perform various experiments. The configuration manual details the architecture of each of the three other models. The architecture of the proposed model is shown in Fig.5 and is described as follows:

- **User layer:** The user index is the input to this layer. The input shape is 1. This is followed by an embedding layer of size 10 and a Flatten layer to convert the matrix to a single array.
- **Product layer:** The product index is the input to this layer. The input shape is 1. This is followed by an embedding layer of size 10 and a Flatten layer.
- **Review layer:** The review layer takes the pre-processed text sequences as inputs. Its input shape is 200. This is followed by an embedding layer to map the word input sequence to a vector representation. An embedding dimension of 128 is used. The embedding layer is passed to an LSTM model with 100 neurons and a Dense layer.
- **Combined model**: The outputs from the user and product layer are concatenated (concatenation 1) and then the output from the LSTM layer, the latent document vector, is concatenated with concatenation 1. This output is passed to three dense layers. The final Dense layer predicts the ratings.

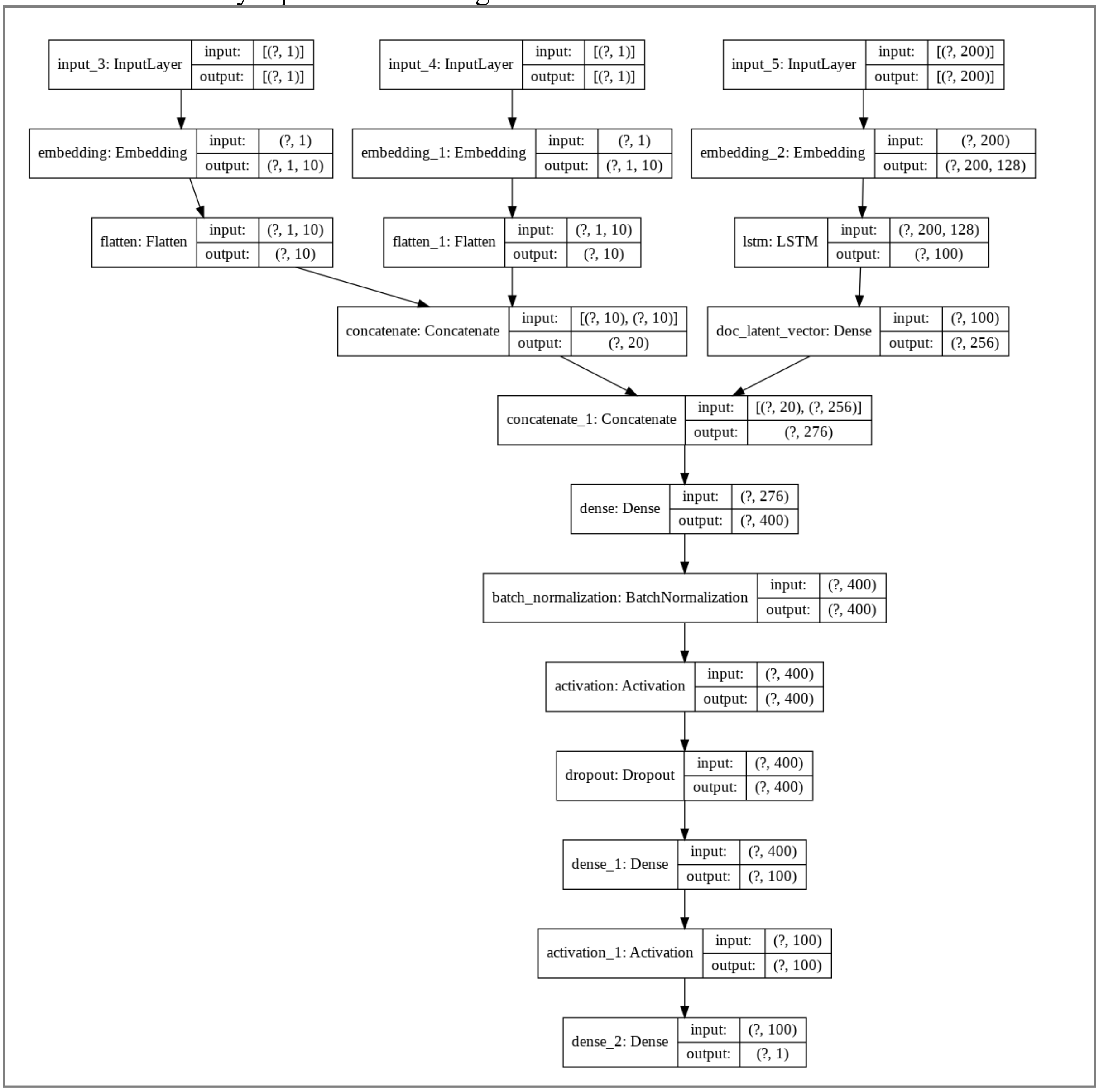

Figure 5: NCF + reviews architecture

4.5 Training

The dataset used was split into two sets. 80% for training and 20% for testing. All the hyperparameters of the proposed methodology and baselines were fine-tuned to achieve optimum efficiency, and the best performance is reported based on the test data. The hyperparameters are: dropout probability is 0.2; the optimiser used in SGD (Stochastic Gradient descent) with a learning rate of 0.001; the loss used is mean squared error; regularisation coefficient is 0.0001. Since the models had different capabilities, the number of epochs needed to train the model accurately varied, so training was carried out at 2, 5 and 10 epochs, respectively.

5 Evaluation

This section presents the experiments carried out to evaluate the effectiveness of each review element in the system. We compare the performance of our model with three other state-of-the-art models. The state-of-the-art models selected for comparisons are explained below. As mentioned in the previous sections, the mean squared error (MSE) was used as the estimation error for the models. Table 3 below compares the results of the three experiments.

Baseline models

- **MF** is the basic and commonly used model for matrix factorisation. It predicts unknown ratings by computing the inner product of the user and item latent factors.
- **HFT** (McAuley and Leskovec, 2013) is an approach that incorporates domain topics learned from reviews into the standard matrix factorisation algorithm.
- **NCF** (He et al. 2017) a deep neural network that models the interaction between users and items instead of traditional matrix factorisation.

In the first experiment, a standard matrix factorisation was used to model the interaction between users and items. The input to the model was the user index and product index to predict ratings. For the second experiment, the LDA topic modelling technique was used to extract the latent topic from reviews, these latent review topics are then modelled simultaneously with the user index and item index for predict ratings. The number of topic distributions used is 5. In the third experiment, a deep neural network was used to model the rating prediction problem using ratings alone. The input to this model was the user index and product index, respectively. The last experiment, which is the core of this research, a deep neural network, was used to model a combination of ratings and reviews. The input to the model is the user index, product index, and text reviews.

Table 3: MSE comparison between the proposed model and other baselines (The lower score, the better).

	MF	**HFT**	**NCF**	**NCF + reviews**
MSE	0.39	0.66	0.32	**0.30**

From Table 2, it is first observed that the neural networks perform better than shallow models which are in line with He et al. (2017). This indicates that using deep learning-based representations of users and items outperforms the vanilla matrix factorisation approach to recommendation systems. Second, comparing our model with HTF (McAuley and Leskovec, 2013), which also models text reviews for recommendations, we see that our model outperforms HFT. This is in line with previous literature findings that demonstrated that deep learning models such as CNN (Zheng et al., 2017) or LSTM (Jin et al., 2018) perform better than topic-based approaches for text processing. The reason for this improvement is because deep learning models have the ability to learn the semantic similarity between terms and phrases, thus, allowing the model to learn more user preferences.

Overall, we find that our proposed model NCF + reviews achieve better performance than other models. This validates our hypothesis that reviews contain more information of user's preferences on these products compared to only ratings and incorporating reviews as side information can enhance rating prediction.

Recommendation

We investigate the result of using the models to make recommendations. First, we select random users considered as a query user. Then we make rating predictions on products for the user in question. Finally, by sorting the rating predictions, we select the top recommended products. Table 4 lists the results of the models for two user queries. Due to space, we only show the three best performing models.

Unsurprisingly, MF, NCF, NCF + reviews reported significant variations between their recommendations. Besides, we check manually and confirm that those recommendations have common characteristics such as Hydration (a desired cosmetic effect). Thus, the results suggest that the models performed well in finding similarities between users and products.

Table 4: List of top recommended products

Query	Rank	MF	NCF	NCF + reviews
User 1	1	Flax	Paula's Choice Hyaluronic acid booster	Rose Face Mask
	2	Virgin Marula Luxury Facial Oil	Sweet almond oil	Macadamia nut oil
	3	Watermelon Glow Sleeping Mask	Beta hydroxy acid	Watermelon Glow Sleeping Mask
	4	Jet Lag Mask	Borage seed oil	rice bran oil
User 6	1	Cerave	Licorice extract	Sugar Face Polish
	2	Cetaphil	Neutrogena Ultra Gentle Hydrating Cleanser	Virgin Marula Luxury Facial Oil
	3	Vitamin C	Protini Polypeptide Cream	rose flower
	4	Hyaluronic acid	Isopropyl palmitate	retinol

5.1 Discussion

This section discusses the meaning, importance and relevance of the findings of this research. It analyses the differences between this research and the previous research on this topic. The section also addresses the research question.

This work investigates how user reviews can be used as a source of skincare recommendation knowledge. We have suggested an approach to represent users and items using natural language processing techniques. The results are promising and indicate that effective opinion mining of text user feedback can be used to understand user preferences. Our model can benefit e-commerce companies by increasing revenue and customer satisfaction.

The results obtained from this research indicated that using deep neural networks to model the primary interaction between users and items outperform the conventional matrix factorisation approaches. This is because deep neural networks can model the non-linearities in the data, as opposed to matrix factorisation models that are linear; hence, they can capture complex user preferences. Though this result is similar to Kiran et al. (2020) where they used embeddings and a deep neural network to learn the non-linear characteristics between items and users, however, this research differs, as the authors only used embeddings and did not take account of the contextual information of long sequences of data such as long comments.

This research shows that joint deep modelling of users and items using reviews can improve rating predictions, which supports existing literature on this topic. One work in close relation to this work is Zheng et al. (2017), where they used two neural networks aligned to learn the latent factors for users and items. One neural network focuses on learning user behaviour, and the other neural network focuses on learning item properties. A shared layer is introduced using matrix factorisation to determine how user and item latent features interact. The approach to this research differs as it is based on a methodology which uses an end-to-end deep learning architecture to model the interaction between users and items. In addition, we consider the underlying sentiment of the review when modelling the user and item representations.

Zhang et al. (2016) have been able to show that using word embedding vector representations can capture more semantic word information compared to bag-of-word vector representations. This is consistent with this research, as this research has also shown that using word embeddings can better learn user preferences compared to related works that used topic distributions. Topic modelling also uses a bag-of-word approach. While Zheng et al. used pre-trained word embedding, this research uses an embedding layer and an additional LSTM layer.

The approach used by Nakajima et al. (2019) to skincare recommendation was to recommend products containing the desired ingredient that has been mentioned in user reviews. Though, there’s no direct comparison between this research and Nakajima's; however, this work helps extend and strengthen Nakajima’s position by showing that building product profiles from text reviews can lead to improvements compared to traditional profiles built with static product details. This research has been able to show that several review elements can be used to improve recommendation. The combination of different review elements such as sentiment score, text reviews, and product indexing is a welcome development as it can be used when explicit user ratings are unavailable or insufficient.

Due to the above results from the experiments performed in Section 5, the aim and objectives of this research, which is to assess how effective user review can be used for recommendations, have been addressed. While the proposed approach is a strong start to a recommendation algorithm, further studies will take into account expanding the data used by adding more feedback and using unsupervised learning techniques to identify products and ingredients. We would also want to expand the skincare dictionary and refine other unsupervised learning methods to automatically pull out the parts of speech and come up with new nouns.

6 Conclusion and Future Work

This research proposed a methodology to skincare recommendation, which has the capability of using text user reviews as the primary source of recommendation knowledge. It is therefore not based on static user-profiles and item catalogues, both of which might not be available in realistic settings. The main components of the system include cleaning of reviews, feature extraction, and using a deep neural network for rating prediction. Based on experimental results, it can be concluded that effective mining of user reviews can reveal some information about customer preferences. By leveraging different review elements for recommendation knowledge, our model is intrinsically more robust to understand the commonalities between users and items.

Furthermore, the proposed system uses an LSTM encoder to better learn the contextual information of words and a deep neural network to model the main interaction between users and items. The results indicate that modelling the interaction of users and items with neural networks outperforms the traditional matrix factorisation approach. This research also demonstrated that the learned user and item latent features are meaningful and effective in providing practical skincare recommendations.

While feature engineering limits the generalizability of the results, this methodology provides new insight into capturing user interest from text reviews to address rating prediction problems. In future, more advanced unsupervised techniques can be leveraged to extract keywords from comments automatically and to perform sentiment analysis.

Finally, this system, deployed in an e-commerce setting can help narrow customer choice, predicting the products most likely to be purchased by the customer, thus increasing business sales & brand loyalty.

References

Alexandridis, G., Tagaris, T., Siolas, G. and Stafylopatis, A. (2019) 'From Free-text User Reviews to Product Recommendation using Paragraph Vectors and Matrix Factorization', In *Companion Proceedings of The 2019 World Wide Web Conference,* May 2019, pp. 335-343, ACM doi: 10.1145/3308560.3316601.

Almahairi, A., Kastner, K., Cho, K. and Courville, A. (2015) 'Learning distributed representations from reviews for collaborative filtering', In *Proceedings of the 9th ACM Conference on Recommender Systems*, September. 2015, pp. 147-154. doi: 10.1145/2792838.2800192.

Catherine, R. and Cohen, W. (2017) 'Transnets: Learning to transform for recommendation', In *Proceedings of the eleventh ACM conference on recommender systems,* August 2017, pp. 288-296, doi:10.1145/3109859.3109878

Chen, L., Chen, G. and Wang, F. (2015) 'Recommender systems based on user reviews: state of the art', *User Modeling and User-Adapted Interaction*, 25(2), pp.99-154. doi: 10.1007/s11257-015-9155-5.

Da'u, A. and Salim, N. (2019) 'Sentiment-aware deep recommender system with neural attention networks', *IEEE Access*, *7*, pp.45472-45484, doi: 10.1109/ACCESS.2019.2907729.

Esparza, S.G., O'Mahony, M.P. and Smyth, B. (2012) 'Mining the real-time web: a novel approach to product recommendation', *Knowledge-Based Systems*, 29, pp.3-11. ScienceDirect. doi: 10.1016/j.knosys.2011.07.007

He, X., Liao, L., Zhang, H., Nie, L., Hu, X. and Chua, T.S. (2017) 'Neural collaborative filtering', In *Proceedings of the 26th international conference on world wide web*, Apri, 2017, pp. 173-182, doi: 10.1145/3038912.3052569.

Jin, M., Luo, X., Zhu, H. and Zhuo, H.H. (2018) 'Combining deep learning and topic modelling for review understanding in context-aware recommendation', In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, June, pp. 1605-1614, doi: 10.18653/v1/N18-1145.

Kim, D., Park, C., Oh, J., Lee, S. and Yu, H. (2016) 'Convolutional matrix factorisation for document context-aware recommendation', In *Proceedings of the 10th ACM conference on recommender systems,* September 2016, pp. 233-240, doi: 10.1145/2959100.2959165.

Kiran, R., Kumar, P. and Bhasker, B. (2020) 'DNNRec: A novel deep learning based hybrid recommender system', *Expert Systems with Applications*, *144*, p.113054, doi: 10.1016/j.eswa.2019.113054

Lee, G. (2020) A Content-based Skincare Product Recommendation System. Department of Computer Science Earlham College.

McAuley, J. and Leskovec, J. (2013) ' Hidden factors and hidden topics: understanding rating dimensions with review text', *In Proceedings of the 7th ACM conference on Recommender* *systems*., *Association for Computing Machinery*, New York, NY, USA, October 2013, pp. 165-172. doi: 10.1145/2507157.2507163.

Nakajima, Y., Honma, H., Aoshima, H., Akiba, T. and Masuyama, S. (2019) 'Recommender System Based on User Evaluations and Cosmetic Ingredients', *In 2019 4th International Conference on Information Technology (InCIT)*, Bangkok, Thailand, 2019, pp. 22-27, doi: 10.1109/INCIT.2019.8912051.

Qiu, L., Gao, S., Cheng, W. and Guo, J. (2016) 'Aspect-based latent factor model by integrating ratings and reviews for recommender system', *Knowledge-Based Systems*, 110, pp.233-243. doi: 10.1016/j.knosys.2016.07.033.

Ramzan, B., Bajwa, I.S., Jamil, N., Amin, R.U., Ramzan, S., Mirza, F. and Sarwar, N. (2019) 'An Intelligent Data Analysis for Recommendation Systems Using Machine Learning', *Scientific Programming*, *2019.* doi: 10.1155/2019/5941096.

Shani, G. and Gunawardana, A. (2011) 'Evaluating recommendation systems', In *Recommender systems handbook,* Springer, Boston, MA, pp. 257-297.

Shoja, B.M. and Tabrizi, N. (2019) 'Customer Reviews Analysis With Deep Neural Networks for E-Commerce Recommender Systems', *IEEE Access*, *7*, pp.119121-119130. doi:10.1109/ACCESS.2019.2937518.

Srifi, M., Oussous, A., Lahcen, A.A. and Mouline, S. (2020) 'Recommender Systems Based on Collaborative Filtering Using Review Texts—A Survey', *Information*, *11*(6), p.317, doi:10.3390/info11060317

Wang, H., Wang, N. and Yeung, D.Y. (2015) 'Collaborative deep learning for recommender systems', In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining,* August, 2015, pp. 1235-1244, doi: 10.1145/2783258.2783273.

Zhang, W., Ding, G., Chen, L., Li, C. and Zhang, C. (2013) 'Generating virtual ratings from Chinese reviews to augment online recommendations', *ACM Transactions on intelligent systems and technology (TIST)*, *4*(1), pp.1-17. doi: 10.1145/2414425.2414434.

Zhang, S., Yao, L., Sun, A. and Tay, Y. (2019) 'Deep learning based recommender system: A survey and new perspectives', *ACM Computing Surveys (CSUR)*, *52*(1), pp.1-38, doi: 10.1145/3285029.

Zhang, Y., Ai, Q., Chen, X. and Croft, W.B. (2017) 'Joint representation learning for top-n recommendation with heterogeneous information sources', In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, November 2017, pp. 1449-1458, doi: 10.1145/3132847.3132892

Zheng, L., Noroozi, V. and Yu, P.S. (2017) 'Joint deep modelling of users and items using reviews for recommendation', *In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, February, pp. 425-434, ArXiv: https://arxiv.org/abs/1701.04783