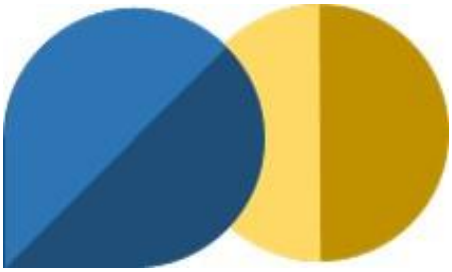
USER PROFILING IN RECOMMENDER SYSTEMS

Deep Learning for Social Media, SS2021, Seminar Paper
Submission Date: 31 August 2021



Dehner, Mirko, Paderborn University, Paderborn, Germany, mideh@campus.upb.de, 7206056
Tenku, Joel Cedric Fotso, Paderborn University, Paderborn, Germany, joelf@campus.upb.de
6810782

Abstract

The recent proliferation of the social media as well the growth of their volume content has increased the need for more and more accurate recommendations. They are based on the analysis of the characteristics of users' interest, which often corresponds to a user profiling. User profiling consists of extracting user labels on different attributes such as age, gender, occupation, income, and interest, which greatly facilitate accurate personalized recommendations. This paper will show the studies about user profiling in recommendation systems and will propose an approach based on the Siamese neural network, which considers users' preferences and item attributes for movie recommendations. Furthermore, the results of this network and others recommendation methods will be discussed.

Keywords: User profiling, Users' preferences, Recommendation Systems, Siamese Neural Network

1 Introduction

Using social media is part of the daily routine of billions of people around the world. According to Johnson (2021), the speed with that the number of internet users around the world grow, is still increasing. Social media as Facebook for social networking; WordPress for blogging; Twitter for microblogging; Flickr and YouTube for photo and video sharing, Digg for social news reading; and Delicious for social bookmarking are the most visited websites in recent years. So, these social medias use information generated by users to create and contribute their content, which will be recommended to other users (Guy et al, 2010). Due to the abundance of information available, users often face the problem of too much choice when looking for content of interest. Thus, companies have created Recommender Systems to guide their users through this.

A Recommender System is a system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options (Wang et al, 2010). Thus, many traditional Recommender Systems such as Popularity-based Recommender Systems, Content-based Recommender Systems and Collaborative Filtering Recommender Systems have been developed to support users. However, these Recommender Systems are not based on User Information and thus will not be able to meet user expectations. To overcome this issue, User Profiling has been integrated in the Recommender System (Kanoje et al, 2015).

User Profiling is today an important task in the building of a Recommendation System. It refers to the extraction of users' demographic information such as age, gender, occupation, interests in order to improve the recommendation of items (Muniyandi and Sheshasaayee, 2021). Due to the advantage to use users' interests and preferences to make recommendations, this paper proposes an approach based on Siamese Neural Networks (SNN) that are using user and item information to recommend relevant items to the users.

In the following project report, we will start by briefly reviewing related research about Recommender Systems, User Profiling and Siamese Neural Networks. In the part of development, we first will describe and explore the movielens dataset, that we will use for the project. Afterwards, we will explain how the SNN has been built and which metrics are used to evaluate the model, and then we will present the results of the model before we conclude with a brief summary and some limitations and outlook.

2 Related Work

2.1 Movie Recommender Systems

Movie Recommender Systems has been today an active area of research and practical applications (Gutti et al., 2016). According to the recommendations, there are many approaches regarding the type of information considered. MOVIEMENDER is a movie recommender system proposed by Gutti et al. (2016) based on the hybrid approach. The hybrid recommender system is the combination between the content based and the collaborative system. The content-based approach is used to compute a pseudo rating matrix, and the collaborative filtering is then applied to this matrix to this matrix to make recommendation for an active user.

Kumar et al. (2015) proposed MOVREC, which is a movie recommendation system based on a Collaborative Filtering approach. Collaborative Filtering recommends movies based on the similarity measures between users and items. It recommends those movies which are preferred by users with similar taste.

2.2 User Profiling

One of the important tasks of building a Recommendation System is analysing the characteristics of users' interests (Muniyandi and Sheshasaayee, 2021). This is often referred to as User Profiling. User Profiling encompasses to the extraction of user labels on different attributes such as age, gender, occupation, income, and interests. Complete and accurate attribute labels will effectively reveal the inherent characteristics of users, thus greatly facilitating accurate personalized recommendations. Therefore, Kanoje et al. (2015) presented a Process of User Profiling: This process includes 3 essential steps: It begins with the profile extraction, which first consists in extracting useful information and documents about the user from Social Network and then in pre-processing information namely segment the text into tokens and assign possible tags to each token. After the profile extraction follows the profile integration. It includes 3 steps namely the creation of the dataset, the removing of various attributes with missing values, and the refinement of the dataset. Finally, the step of interest discovery as the last step of the process of user profiling, which consists in discovering the latent topic distribution associated with each user. Hung et al. (2008) proposed a new approach to user profiling based on the tags associated with the user's personal collection of social media in order to explore the role of tagging for Social Media Recommendation. In their study, they used a popular social bookmarking website "Del.icio.us", which has rich and public personal bookmark collection. Each user from the dataset is profiled by aggregating the tags specified by the user as well as his/her social contacts. Unlike many traditional recommender systems, which based on the relationship between user preference and item attribute from rating data, their proposed approach is based on the relationship between user attributes (user tags) and content attributes in order to determine recommendation score. Therefore, all items with recommendation scores higher than threshold can be recommended to users.

Tutubalina and Nikolenko (2017) presented a Convolutional Neural Network (CNN) regarding medical application in order to extract or predict demographic information of users. To do that, they have collected a database of medical reviews from a health-related website with User-Generated Content (UGC), namely WebMD, and then have applied on this dataset topic models with user attributes such as PLDA and USTM, and neural models based on top of word2vec embeddings. And finally, they have trained these models to predict the age and gender of users who wrote these reviews.

2.3 Siamese Networks

In order to predict recommendations for a single user on the basic of the user data such as age, gender, preference, Banjac et al. (2020) prosed the Movinder recommendation system based on the SNN to predict recommendations for a group of users. For that, they used the MovieLens dataset, which contains 100,000 ratings from 943 users on 1,682 movies. The original idea is building Bilinear Neural Network with Ranking Loss (Triplet Loss) and combine them into a Siamese Network Architecture.

Using the input data, the SNN creates embeddings and generates vector outputs that show the actual distance between movies in terms of their complementary. In the SNN, the inputs are user, positively rated items, and negatively rated items with the same weights, which give as outputs the single embedding for the items. For their project, they obtained a 94.5% accuracy on the train and 91.1% of accuracy on the test set. Therefore, they concluded that the Siamese neural network is appropriate to recommend movies to new users.

3 Development

3.1 Dataset

To train the model, we use the publicly available dataset MovieLens[1] on movie ratings and movie characteristics. MovieLens dataset was collected by GroupLens[2] Research Project at the University of Minnesota. From these MovieLens datasets, we use specifically the Movielens-1M dataset[3], which contains 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users. Ratings range between 1 and 5, where 1 corresponds to a very bad note and 5 to a very good note. Each user in the dataset has rated at least 20 movies. On average, each user has rated about 166 movies (cf. scatterplot in Figure 1), which makes 4.47% of the listed movies, resulting in a sparse user-item-interaction matrix.

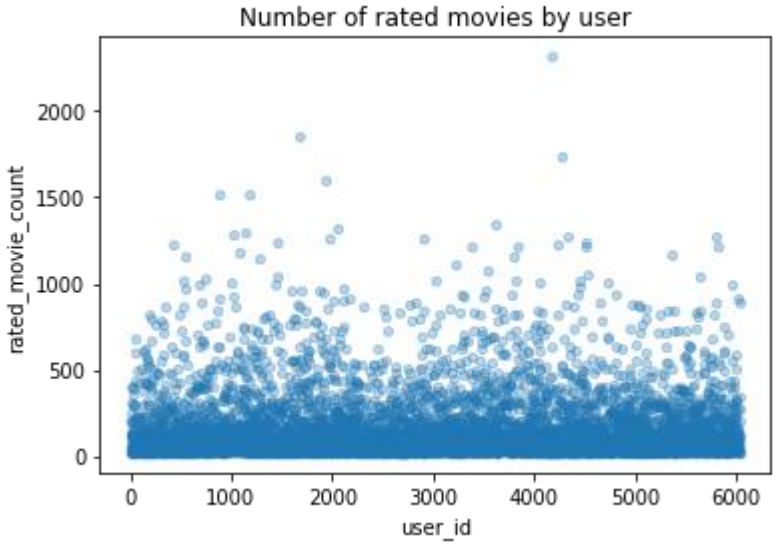


Figure 1: Number of movies being rated by users

User entries also include demographic information, specifically the age, the gender, the occupation, and the zip code. Gender is denoted by a "M" for male and "F" for female.

The dataset provider has also annotated each movie with a list of genres. There are in total 18 genres such as action, adventure, animation, comedy, crime, drama, horror, etc. Each of the movies also have a composite text that concatenates the name (the title) with the year of release.

[1] http://www.movielens.org/

[2] http://www.grouplens.org/

[3] https://grouplens.org/datasets/movielens/1m/

When further exploring the dataset, you can see some interesting characteristics:

According to the timestamps present in the dataset, the vast majority of ratings in the dataset were created in the year 2000. You can see the distribution in the following Figure 2.

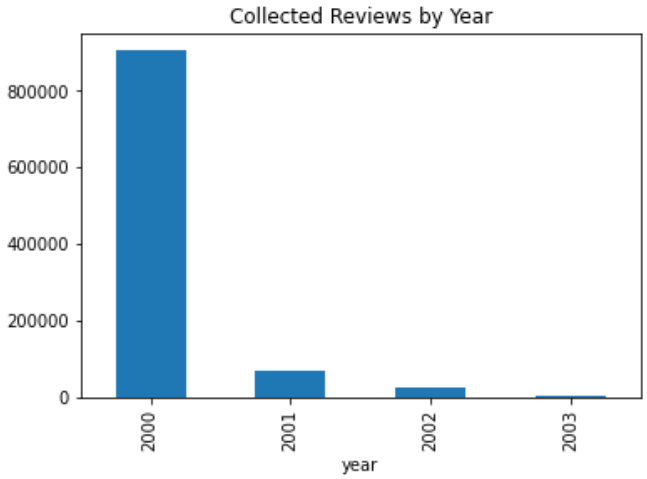


Figure 2: Movie Ratings by Year

The following Figure 3 shows how often movies from each year have been rated (red) and the amount of movies that have been released in that year (blue). You can see that there are not many really old, classic movies in the dataset – most of the movies are younger than 20 years. You can also see that the ratings for movies between the years 1975 and 1990 are a bit overrepresented as the proportion of ratings to released movies is clearly higher than in the other years. (In other words: There are more ratings per released movie than in the other years.) It is not expected that this will affect the model performance though.

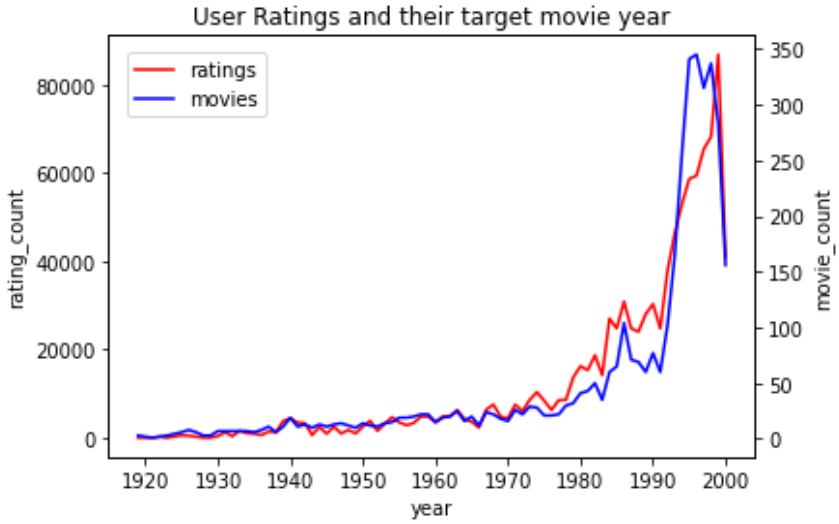


Figure 3: Ratings and Released Movies by Year

As ratings only have discrete values between 1 and 5, the boxplot for the given ratings in Figure 4 shows only some information. The mean rating is 4.0 – the mark in the boxplot is obstructed by the line of the upper quantile. It shows that overall, users tend to give high ratings as you can see in the histogram in Figure 5.

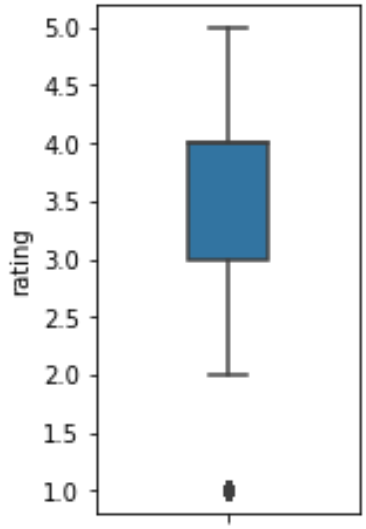


Figure 4: Ratings Boxplot

Taking a closer look at the distribution of ratings, you can see a slightly shifted, bell shaved curve. This is remarkable, because usually online reviews tend to "[…] follow an asymmetric bimodal (J-shaped) distribution with more extreme positive (5-stars) than extreme negative (1-star) reviews on a 5-star Likert-type scale" (Hu, Pavlou et al., 2007, p. 2).

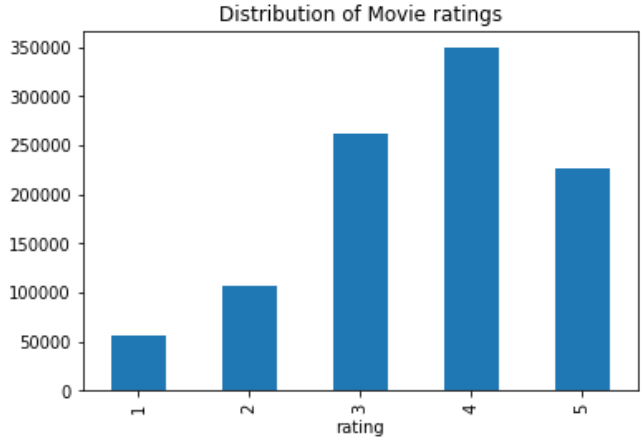


Figure 5: Ratings Histogram

The correlation plot in Figure 6 shows that there are no relevant correlations between the numerical features.

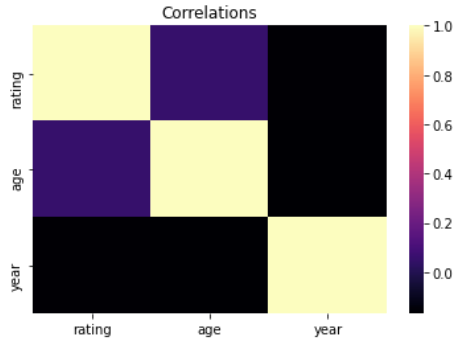


Figure 6: User Feature Correlations

In the next correlation in Figure 7, you can see the correlations between genres. Additionally, there are also the correlations with the rating shown in the last row. It shows that for example there are many movies that are associated with Action and Adventure, while there are only rare examples where a movie is an Action and Drama movie.

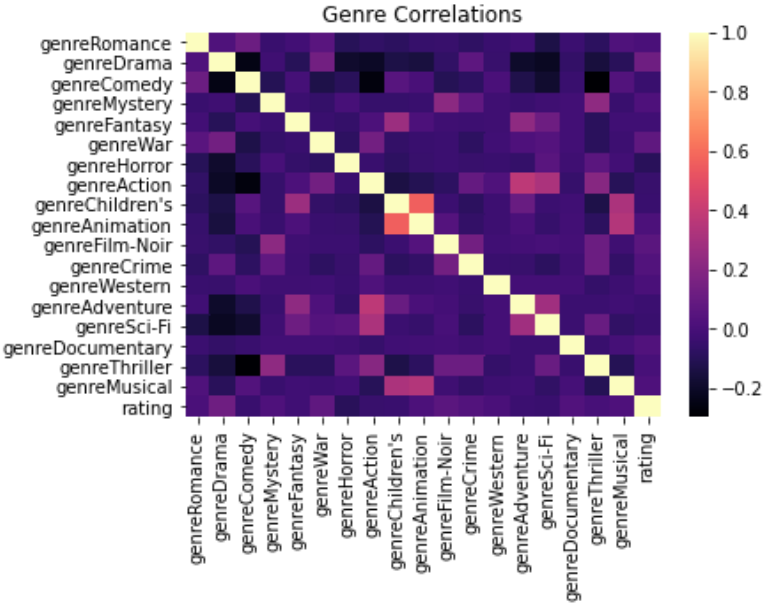


Figure 7: Movie Feature Correlations

3.2 Model

The given dataset is finally used to develop a movie recommender system based on a siamese neural network. In order to make the MovieLens dataset feasible for this approach, certain pre-processing steps had to be taken before implementing the model.

First, duplicates were removed from the dataset, using the User ID for the file “Users” and Movie ID for the file “Movies” as key identifiers. After the removal was executed, the dataset contained 6,040 users rating 3,007 movies. The number of ratings stored in the file “Ratings” remained unchanged, yielding a total sum of 1,000,029 ratings. Furthermore, the movie genre was encoded into a dummy variable in order to enable information gain through data exploration. For that, the existing movie genre column was split into a set of all available movie genres, in which each genre has been assigned to a separate column. After doing so, the encoding of the movie genres into dummy variables was executed and added to the DataFrame. Furthermore, the user ID, gender, age and occupation, which are relevant for the following neural network has been stored as a separate variable, whereas the irrelevant ZIP-code has been dropped. The categorical variable Gender had to be transformed into a numerical variable with female as 0 and male as 1 in order to work as a feature input for the embeddings (cf. Table 1). Likewise, the values for age and occupation have been categorized into numerical values (cf. Table 2 and Table 3). For implementing the SNN, the creation of a user-interaction matrix was required, where pre-processed data from the three given files “movies”, “users” and “ratings” had to be merged. Therefore, the ratings were matched to their corresponding movies and their users based on their unique ID. Since, in the user-interaction matrix the previously used movie and user IDs are no unique identifiers anymore, new consecutive movie und user ID’s have been computed and used as new key identifiers in the following process.

Our SNN was implemented by using *keras* in Python. The user age, its gender, occupation and submitted movie ratings were used as input variables for the model. After a final drop of duplicates in the

input variables, the number of user features remained unchanged at 6,040 and 3,705 as movie features. For model evaluation we use holdout validation. Therefore, the dataset was split into training and test data containing the new user and movie IDs as key identifiers, the corresponding ratings and the rating timestamp. Finally, eighty percent of the data were assigned to the training set and twenty percent to the test set. As movie recommendations to users are the expected output of the model, a determination about the ratings is required, concerning the question which movie rating represents a movie the user liked and which rating does not. For the following model, the assumption is made that only movies with a high rating depict movies the user liked. Therefore, it has been determined that movies with ratings with a value of 4 and 5 were defined as movies the user liked, whereas movie ratings below 4 were considered as movies the user did not like. Hence, movies ratings with values of 4 and 5 depict a recommendation by the user to watch the movie. Based on that, the ratings were transformed into a binary variable accordingly.

For the model, a triplet-loss function has been computed. This loss-function is depicted by three inputs from the given dataset. First, the anchor is represented by the variable user input, which contains the user ID. Second, the positive input is represented by the variable containing the movies the user liked. Lastly, the negative input is represented by the variable containing movies the user did not like. Finally, the goal is pursued to minimize the distance between the anchor and the true input on the one hand and maximize the distance between anchor and negative input on the other hand. (Dong, Shen 2018). Additionally, a Bayesian Personalized Ranking Loss (BPR) was computed as a feasible method to solve the personalized ranking task (Rendle, et.al. 2009).

$$p(i >_u j|\Theta) := \sigma(\hat{x}_{uij}(\Theta))$$

where σ is the logistic sigmoid:

$$\sigma(x) := \frac{1}{1 + e^{-x}}$$

…..

The shape of a is (batch_size, n_user_features), p and n are both arrays of shape (batch_size, n_item_features). Each row of the array is represented by a feature vector.

First, we learn Embeddings for user and movie feature inputs. In the typical Siamese Network approach, we learn multiple (“Siamese”) layers to perform different, but similar tasks in parallel (cf. Section “Related Work”). In the end, the results of the Siamese Layers are fed into the TripletLoss-Layer that keeps track of the deviations between anchor, positive and negative examples.

Finally, our Siamese Neural Network is based on the following Architecture in Figure 8:

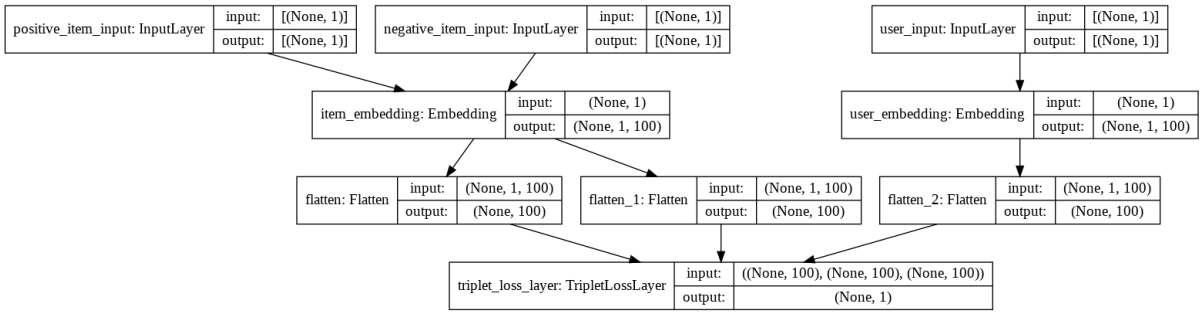


Figure 8: Siamese Neural Network Architecture

As you can see, we decided to use a latent dimension of 100 for the Embedding Layers. This is a Hyperparameter you can set. Increasing the dimensionality gives slightly better results but comes at a

price of increased training time. We did not perform extensive Hyperparameter Tuning at this point because computing resources were limited.

3.3 Evaluation

The reported training losses and accuracies from Figure 9 state that training progresses successfully and converges to an accuracy of 92.5% on the test set. After 15 epochs, the model is not expected to gain better results by further training.

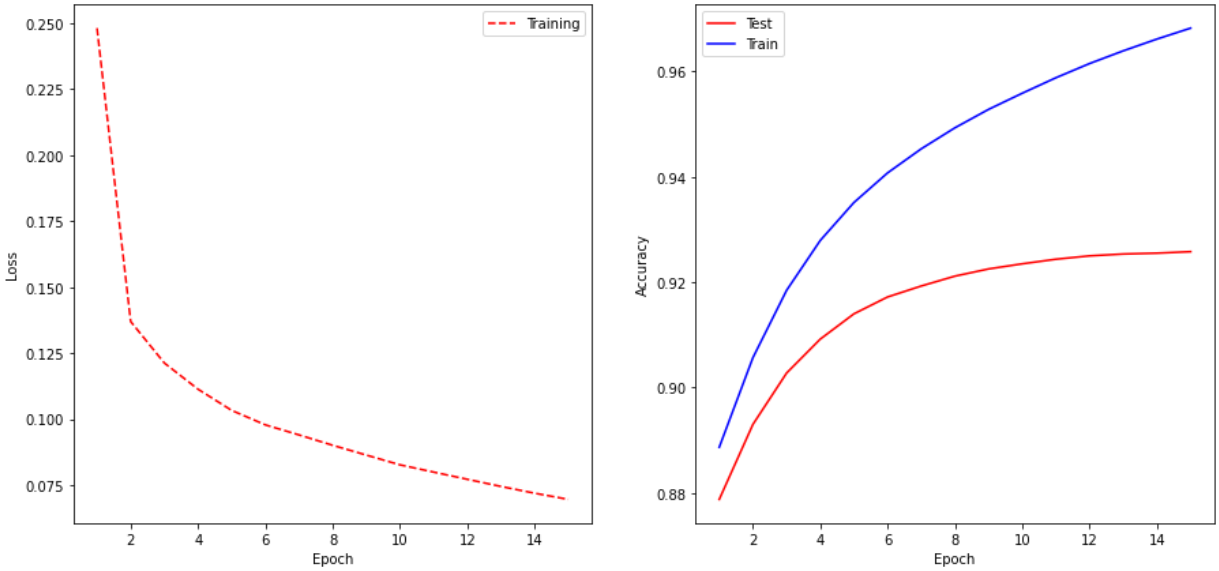


Figure 9: Training and Test Reports

For evaluation of the SNN performance, we compare the results against a baseline. We decided to use the AutoEncoder model from the lecture. It needed some slight modifications to fit into our setup and to run on the Movielens-1M dataset. Both models received the same split of test and training data for best comparability.

In Figure 10, you can see the loss during the training phase of our baseline constructed by AutoEncoders. We trained it for 200 epochs. You can see that it yields a much higher loss than the SNN approach in Figure 9.

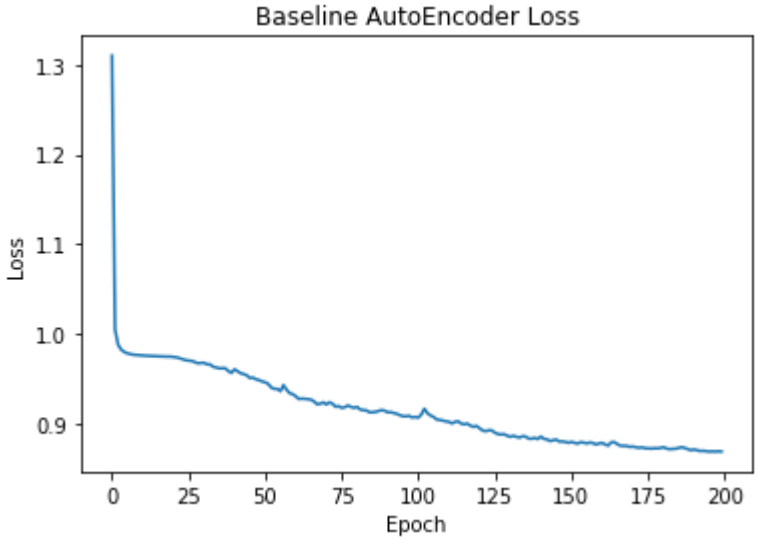


Figure 10: Loss during Training of the Baseline

4 Conclusion

User profiles have been represented by vectors of attributes such as age, name, gender, occupation, affiliation interest of each user. These profiles are the result of the process of User Profiling and plays a fundamental role in the recommendation. User Profiling consists of extracting information about the user from a large amount of structured and unstructured data using topics modelling such as PLDA and USTM, and neural models based on top of word2vec embeddings. In this paper, we have presented an approach based on SNN for Movie Recommendations. From a given set of user attributes, the network selects and proposes a list of relevant movies to a user. Positively and negatively rated movies are the inputs of the network. The process starts by binarizing the user ratings. To compute the loss of the network, we defined a triplet loss layer function, which is the difference between one and the sum of positive and negative items weighted by the user and the identity loss, which represents the difference between true and predicted values. The loss is thus calculated from the user inputs embedded and then used to train the network. After that the network has been trained on 80% of the dataset and tested on the rest. After 15 epochs the training accuracy archives 90% and test 92.5%. Comparing to other studies based on the user profiling such as Tag-Based User Profiling for Social Media Recommendation which used tag to tag matrix to make recommendations (Hung et al., 2008) with 20.3% as accuracy for recommendations for new users, the items being recommended by a SNN are more accurately.

This project has shown that SNN can perform as a Recommender System quite well despite it was designed to work as a Computer Vision approach. Also, we showed that User Profiling can be a good measure in Recommender Systems to tackle the cold-start problem. Further work can focus on hyperparameter tuning as this work did not have the computing resources to do that. Regarding the datasets being used, we came to the result that the way the MovieLens dataset is constructed is more or less outdated. It would rather make sense to provide an updated version with more latent, behavioral features as it is highly unlikely that you are able to collect that kind of demographical data about the users before they are interacting with the system for the first time. This results in a limited applicability of this work for practical contexts. Another emerging research topic in Recommender Systems are graph-based methods, which are similarity-based approaches.

5 References

Banjac, C. Y. Altinigne, S. Kypraiou, P. Sioulas (2020). “Movinder: A Movie Recommendation System for Groups “. EPFL, January 2020, Lausanne, Switzerland.

Dong, X., Shen, J. (2018) “Triplet Loss in Siamese Framework for Object Tracking”, Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, China

Guy, N. Zwerdling, I. Ronen, D. Carmel, E. Uziel (2010). " Social Media Recommendation based on People and Tags “. SIGIR’10, July 19–23, 2010, Geneva, Switzerland. Copyright 2010 ACM 978-1-60558-896-4/10/07.

Hande , A. Gutti, K. Shah, J. Gandhi, V. Kamtikar (2016).” MOVIEMENDER- A MOVIE RECOMMENDER SYSTEM”. International Journal of Innovations in Engineering and Technology (IJIET)Volume 5 Issue 11, November 2016.

Hu, Nan & Pavlou, Paul & Zhang, Jie. (2007). “Why Do Online Product Reviews Have a J-Shaped Distribution? Overcoming Biases in Online Word-of-Mouth Communication”. SSRN Electronic Journal. 10.2139/ssrn.2380298.

Hung Chia-Chuan, Yi-Ching, Huang, Jane Yung-jen Hsu and David Kuan-Chun Wu (2008). “Tag-Based User Profiling for Social Media Recommendation “

Kanoje, Sumitkumar, S. Girase, and D. Mukhopadhyay (2015). "User profiling trends, techniques and applications."

Manoj, D.K Yadav, A. Singh, V. Kr. Gupta (2015). “A Movie Recommender System: MOVREC”. International Journal of Computer Applications (0975 – 8887) Volume 124

Muniyandi, A. Sheshasaayee (2021). “An Evaluation on User Profiling Methodologies and the Challenges Associated with it in Recommender Systems”. Volume 17, Issue 1, January - 2021 ISSN: 1007-1172. Journal of Shanghai Jiaotong University.

Rendle,S., Freudenthaler, C., Gantner, Z. and Schmidt-Thieme, L. (2009). “BPR: Bayesian Personalized Ranking from Implicit Feedback” Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. 453-457

Statista (2021). Internet users in the world 2020 | Statista. Available online at https://www.statista.com/statistics/617136/digital-population-worldwide/ (accessed 2/27/2021).

Statista (2021). Most used social media 2020 | Statista. Available online at https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed 2/27/2021).

Tutubalina, S. Nikolenko (2017). Exploring convolutional neural networks and topic models for user profiling from drug reviews. Springer Science+Business Media, LLC 2017. https://doi.org/10.1007/s11042-017-5336-z

Wang, Y. Tan, M. Zhang (2010). " Graph-based Recommendation on Social Networks “. 12th International Asia-Pacific Web Conference. School of Electronics Engineering and Computer Science. Peking University

APPENDICES

Appendix A – Feature Mappings

F	0
M	1

Table 1: Gender Mappings

Under 18	1
18-24	18
25-34	25
35-44	35
45-49	45
50-55	50
56 and more	56

Table 2: Age Mappings

Other or not specified	0
Academic or educator	1
artist	2
clerical or admin	3
college or grad student	4
customer service	5
Doctor or health care	6
Executive or managerial	7
farmer	8
homemaker	9
k-12 student	10
lawyer	11
programmer	12
retired	13
Sales or marketing	14
scientist	15

self-employed	16
Technician or engineer	17
Tradesman or craftsman	18
unemployed	19
writer	20

Table 3: Occupation Mappings