Sentiment Analysis on Netflix

PRITI BAGKAR[1], AISHWARYA BORUDE[2], ZARRIN AGA[3]
[1, 2, 3] *Student, Department of Information Technology, Shah & Anchor Kutchhi Engineering College, Mumbai, India.*

Abstract- ***We use machine learning to build a personalized movie scoring and recommendation system based on the user's previous movie ratings. Different people have different taste in movies, and this is not reflected in a single score that we see when we Google a movie. Our movie scoring system helps users instantly discover movies to their liking, regardless of how distinct their tastes may be. Current recommender systems generally fall into two categories: content-based filtering and collaborative filtering.***

We experiment with both approaches in our project. For content-based filtering, we take movie features such as review and keywords as inputs and use TF-IDF and doc2vec to calculate the similarity between movies. For collaborative filtering, the input to our algorithm is the observed users' movie rating, and we use K-nearest neighbors and matrix factorization to predict user's movie ratings. We found that collaborative filtering performs better than content-based filtering in terms of prediction error and computation time.

Indexed Terms- ***NLU, LSTM, Neural Network, Recommendation***

I. INTRODUCTION

Nowadays, many people want to watch TV-shows or -series anytime and anywhere they want. In recent years, online TV has experienced exponential growth. To be exact, regarding the Digital Democracy Survey by Deloitte, which is an annual survey about changes in the digital environment, 49% of the United States households subscribed to one or more streaming video services in 2016, compared to 31% in 2012. An interesting aspect of this exponential growth is the difference in age and the way people watch TV-shows. As can be seen in Fig. 1, there is a big difference between the millennials age between 14 and 31 and the seniors age of 68 + regarding watch behaviour. The millennials prefer not to watch on TV only anymore, as seniors watch on TV almost all the time. Instead, the millennials often choose a mobile device.

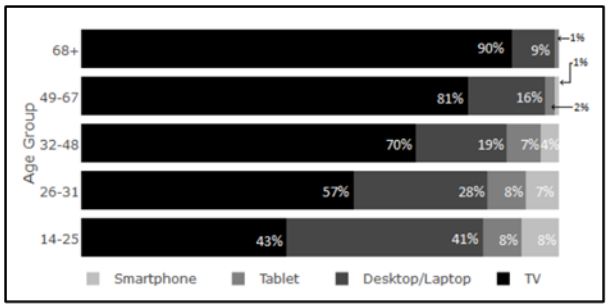

Fig 1. Share of time spent watching TV-shows per device and age group.

Besides the time management advantages that online TV brings to people, another reason people often choose for online and on-demand TV is the absence of commercial breaks, the ability to watch where they want, on which device they want, and the ease to discover new content. Netflix is one of the parties that jumped into the world of online streaming services. Netflix, which was founded in 1999 as an online video shop, has become the most used, and still a strong growing American online streaming provider specialized in video-on-demand distribution. Currently, they are active in over 190 countries all over the world with over 100 million subscriptions. Recently, the number of Netflix subscriptions within the United States exceeded the number of subscriptions for regular paid cable TV, see Fig. 2. Every day, over 125 million hours of video is watched on Netflix, and the number of titles keeps increasing.

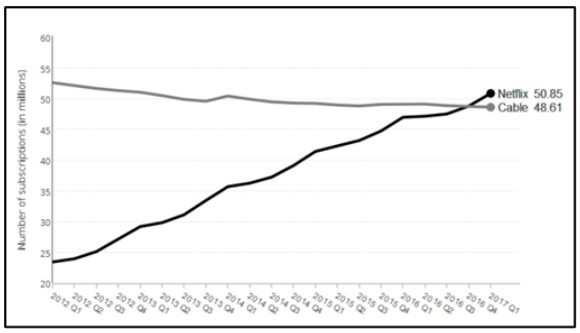

Fig 2. Number of subscriptions in the United States.

From these numbers, one can conclude that Netflix collects a lot of data, which can be used in many ways. For example, they can analyze data to increase the revenue, for marketing purposes, and to improve their customer satisfaction. Regarding customer satisfaction in general, recommendations based on the user's behaviour has played an important role in e-commerce customer satisfaction. Many web shops, like Amazon and Alibaba, use recommender techniques to recommend items to their visitors, which are items that are similar to the one they searched for, or they have bought recently. Next to that, recommender systems are also widely used by online travel agencies like booking.com and Expedia, so that visitors can discover their ultimate holiday destination match, based on their search behaviour, historic bookings, or similar users. In fact, recommender systems are used in all kinds of industries due to the enormous data-driven environment we are living in nowadays. Besides the increasing number of social media platforms, which all generate a tremendous amount of data, the need of users to personalize content has also played an important role in the development of recommender systems. Not only is Netflix using recommender systems to improve customer satisfaction, but also because people are bad at choosing between many options. From consumer research Netflix has conducted, it suggested that an ordinary Netflix user loses interest after 60 seconds of choosing or reviewing more than 10 to 20 titles in detail. Therefore, Netflix developed a recommender system over the years, which consists of various algorithms that are combined into an ensemble method.

It seems obvious that one could state a recommender system is vital for a company such as Netflix. Therefore, optimizing and fine-tuning these kinds of models will tremendously increase the customer satisfaction and therefore the overall revenue. This paper will cover the process of building a recommender system from start to finish. Therefore, the research question of this paper is: Which recommender technique applied to Netflix movie data will perform best? And will the extension of additional data improve this model? First, we will discuss recent literature that has been conducted in the field of recommender systems. Next, the data that is used to train the models will be pre-processed and analyzed. Thereafter, the actual recommender systems will be trained. In this research, we will use customer ratings only at first. Later on, it will be extended with external metadata, such as actors, genres, IMDb ratings, and release dates. Finally, several evaluation methods will be applied to the models, and one final recommender system will be advised.

II. BACKGROUND

Since recommender systems are such a hot topic in recent data science research, many scientific articles have been published in the field of recommender systems. In this section, we will discuss several relevant works that have been published. Recommender systems can be roughly divided into three groups: collaborative filtering, content-based filtering, and hybrid filtering. Collaborative filtering is a recommender technique that focuses on the interest of the user, by using preferences of other similar users. The psychology behind this approach is that if user 1 and user 2 can be considered as having the same interests, one can assume user 1 has also the same opinion about a new item only user 2 has already an opinion of.

Sarwar et al. (2001) divide collaborative filtering into two categories: memory based collaborative filtering algorithms and model-based collaborative filtering algorithms. Memory-based algorithms use all available user-item data to generate a prediction.

Based on all data it determines the most related users, similar to the target user. These neighbours are similar because they have statistically common interests. To determine these so-called neighbours, several statistical techniques are used. Finally, the top n most

similar items are recommended for the target user. The memory-based collaborative filtering algorithms are also called user-based collaborative filtering algorithms. The advantage of user-based collaborative filtering is the sparsity and scalability. Many recommender systems use data with lots of users and items, but with relatively few actual ratings. User-based collaborative filtering only uses necessary data, which reduces the run time.

Model-based collaborative filtering first builds a model of user ratings only. To do this, it uses several machine learning techniques, such as clustering, rule-based and Bayesian network approaches. Each of the machine learning techniques uses its own approach. The clustering model formulates collaborative filtering as a classification problem, while the Bayesian network model treats it as a probabilistic model and the rule-based model as an association-rule model. The model-based collaborative filtering algorithms are also called item-based collaborative filtering algorithms.

Next to collaborative filtering, one is also able to build recommender systems by using the content of items, and a profile matched to items. This approach is called content-based filtering. Lops et al. (2011) stated that the recommendation process of a content-based recommender system basically consists of matching the attributes of a user profile against the attributes of a content object. The outcome of this process is just the level of the user's interest in an object. It is crucial for a content-based model that the user profile is accurate.

A weakness of collaborative and content-based filtering mentioned by Lika et al. (2014) is the problem of handling new users or items. Both techniques mentioned before are based on historic data of the users or items. This well-known problem is often called 7 the cold-start problem. Burke (2007) suggests hybrid systems might resolve the cold-start problem. In many fields in data science, different kinds of approaches are combined to come to the best result. This process of combining multiple algorithms into one algorithm is often called an ensemble. In the area of recommender systems, a common ensemble method is called hybrid filtering. According to Burke (2007), hybrid recommender systems are any kind of recommender system that combines multiple recommendation techniques to produce output. Therefore, Burke (2002) proposes that a collaborative filtering system and a content-based filtering system can ensemble into one on several ways:

- Weighted: each recommender system in the ensemble has a weight and a numerical combination is made for the final model.
- Switching: the final recommender system chooses a recommender system in the ensemble and applies the selected one.
- Mixed: a combination of different recommender systems is made.
- Feature combination: different data sources are used to gather information and are used in one recommender system.
- Feature augmentation: multiple recommender systems are applied after each other such that the output of each recommender system creates a feature that is used as input for the next recommender system.
- Cascade: there is a strict order in different recommender systems, where the order is chosen such that the weak recommender does not overrule the stronger one. The methodology behind this approach is that the weak recommender can only refine the stronger recommender.
- Meta-level: this technique is in some way equal to the feature augmentation technique. However, the difference between these techniques is that the metalevel approaches produce a model instead of a feature as output. Next, this model is used by another recommender within the ensemble.

III. METHODOLOGY

- Neural network

A neural network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes. Thus, a neural network is either a biological neural network, made up of real biological neurons, or an artificial neural network, for solving artificial intelligence (AI) problems. The connections of the biological neuron are modeled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a

linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic synapses and other connections are possible. Apart from the electrical signaling, there are other forms of signaling that arise from neurotransmitter diffusion.

Artificial intelligence, cognitive modeling, and neural networks are information processing paradigms inspired by the way biological neural systems process data. Artificial intelligence and cognitive modeling try to simulate some properties of biological neural networks. In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents in computer and video games or autonomous robots.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

Neural network theory has served both to better identify how the neurons in the brain function and to provide the basis for efforts to create artificial intelligence.

- Artificial intelligence

A neural network (NN), in the case of artificial neurons called artificial neural network (ANN) or simulated neural network (SNN), is an interconnected group of natural or artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network.

In more practical terms neural networks are non-linear statistical data modeling or decision-making tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data.

An artificial neural network involves a network of simple processing elements and artificial neurons which can exhibit complex global behavior, determined by the connections between the processing elements and element parameters. Artificial neurons were first proposed in 1943 by Warren McCulloch, a neurophysiologist, and Walter Pitts, a logician, who first collaborated at the University of Chicago. One classical type of artificial neural network is the recurrent Hopfield network.

The utility of artificial neural network models lies in the fact that they can be used to infer a function from observations and also to use it. Unsupervised neural networks can also be used to learn representations of the input that capture the salient characteristics of the input distribution, e.g., see the Boltzmann machine (1983), and more recently, deep learning algorithms, which can implicitly learn the distribution function of the observed data. Learning in neural networks is particularly useful in applications where the complexity of the data or task makes the design of such functions by hand impractical.

- Long short-term memory

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected

handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs.

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

The advantage of an LSTM cell compared to a common recurrent unit is its cell memory unit. The cell vector has the ability to encapsulate the notion of forgetting part of its previously stored memory, as well as to add part of the new information. To illustrate this, one has to inspect the equations of the cell and the way it processes sequences under the hood.

IV. IMPLEMENTATION

Using the dataset archived by netflix we are able to train our model using this set. First, we separated and labeled all the reviews as negative or positive then we reshuffled them to create a train test dataset. This dataset is converted from word to vector from so that a computer option can be performed on it. then we are spling the data into testing and training datasets. Here we have kept the ratio of 0.25 which indicates training of 75% of the dataset where 25% of data will be used for testing purposes.

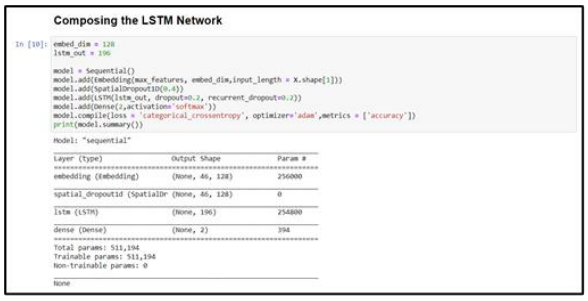

Fig 3. LSTM Neural network.

As shown in fig 3 in the code block, we are building our neural network and sequential input layer whereas and dence 2 node output layers. We are using one Spial drop layer and one LSTM layer as hidden layers between input and output nodes.

Post the creation of ML neural network using code block show in fig 4 we are running the test and epoch is executing. After two epochs we are getting the accuracy of 77% and each epoch takes around 22 seconds to 5 minutes to complete one epoch cycle.

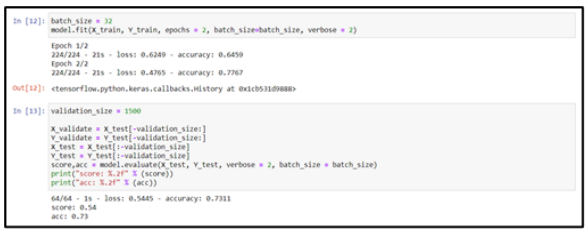

Fig 4. LSTM result analysis.

CONCLUSION

For content-based filtering, we use the movie similarity matrices generated by TF-IDF and word2vec to predict a user's movie ratings. To combine the two similarity matrices from TF-IDF, we calculate their weighted sum. Evaluating the performance on a training set consisting of 80% randomly selected user-movie rating pairs, we see that the RMSE is smallest for the weight w1 = 0:7;w2 = 0:3 as shown in gure 1. We then run the prediction algorithm with these weights on the test set. the performance of our algorithm in predicting users' ratings for movies. Each bin represents a user-movie pair that has a rating in the specific range. The blue bars represent the portion for which our algorithm's prediction is within 0:75 of the true rating, and the green bars represent the portion for which our

algorithm's prediction is outside of 0:75 of the true rating. We see that our algorithm performs well for user-movie pairs with ratings higher than 3, which constitute the majority of the data points. The RMSE on the test set is 1.052. In the next semester we will be working on UI and implementation of our ML model to create a web portal which recommends movies based on collaborative filtering approach which we have made in this semester.

ACKNOWLEDGMENT

We are grateful to the management of Shah & Anchor Kutchhi Engineering College for providing us the facility for the completion of our task. Firstly, we extend our gratitude to Dr. Bhavesh Patel, Principal of Shah and Anchor Kutchhi Engineering College for his continuous support. We express our heartfelt thanks to our guide Prof. T P Vinutha for her valuable guidance and advice related to this work.

Our note of thanks goes to our most cherished Head of Department Prof. T P Vinutha for her undiminished trust and her support throughout our tenure and for giving us the opportunity to work on our project which made us capable of handling assignments on our own and becoming relevant.

We thank all faculty members of the Department of Electronics and Telecommunication Engineering for all the help extended to us and for motivating us. We also extend our gratitude to technical staff in the lab, for all their support and help. On this occasion, we remember the valuable support and prayers offered by our parents, members and friends which were indispensable for the success.

REFERENCES

[1] Deloitte US, “Deloitte Digital Democracy Survey, 11th edition”

[2] Deloitte US, “Deloitte Digital Democracy Survey, 9th edition”, URL: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Technology-Media-Telecommunications/gx-tmt-deloitte-democracy-survey.pdf

[3] Netflix Investor Relations, “Netflix Company Profile”, URL: https://ir.netflix.com

[4] Schwartz B. (2015), “The Paradox of Choice: Why More Is Less”. Harper Perennial, New York, NY.

[5] Sarwar et al. (2001), “Item-Based Collaborative Filtering Recommendation Algorithms”.

[6] Lops et al. (2014), “Content-based Recommender Systems: State of the Art and Trends”