IJWR

A Movie Recommender System Based on Topic Modeling using Machine Learning Methods

Mojtaba Kordabadi[a], Amin Nazari[b], Muharram Mansoorizadeh*[c]

[a] MSc, Computer Engineering Department, Bu-Ali Sina University, Hamedan, Iran; m.kordabadi@gmail.com
[b] Ph.D. Candidate, Computer Engineering Department, Bu-Ali Sina University, Hamedan, Iran; aminnazari91@gmail.com
[c] Associate Professor, Computer Engineering Department, Bu-Ali Sina University, Hamedan, Iran; mansoorm@basu.ac.ir

ABSTRACT

In recent years, we have seen an increase in the production of films in a variety of categories and genres. Many of these products contain concepts that are inappropriate for children and adolescents. Hence, parents are concerned that their children may be exposed to these products. As a result, a smart recommendation system that provides appropriate movies based on the user's age range could be a useful tool for parents. Existing movie recommender systems use quantitative factors and metadata that lead to less attention being paid to the content of the movies. This research is motivated by the need to extract movie features using information retrieval methods in order to provide effective suggestions. The goal of this study is to propose a movie recommender system based on topic modeling and text-based age ratings. The proposed method uses latent Dirichlet allocation (LDA) modelling to identify hidden associations between words, document topics, and the levels of expression of each topic in each document. Machine learning models are then used to recommend age-appropriate movies. It has been demonstrated that the proposed method can determine the user's age and recommend movies based on the user's age with 93% accuracy, which is highly satisfactory.

Keywords: ***Recommendation Systems, Text Classification, Topic Modeling.***

1. Introduction

Families often enjoy watching movies together. Many parents are concerned about selecting the appropriate movies for their children. As a result of the modern and busy lifestyle, it takes a lot of time for people to find and choose their favorite movies. Therefore, they check various programs and websites to find movies with suitable content for their children and grandchildren. Despite the increasing volume of movie and television productions, it is very difficult to find appropriate content in this vast amount of information [1]. Text retrieval techniques have been widely used in many different areas such as web retrieval [2], image and video retrieval [3], and content-based recommendation [4]. Web-based recommender systems are increasingly being used for practical purposes such as entertainment, education, and e-commerce. As recommender systems use similar methods to identify relevant data, they have been closely linked to information retrieval systems. Valuable information can be discovered and extracted from movies using information retrieval models that can also be employed in recommendation systems. Thus, the presence of an efficient recommendation system that uses intelligent methods to display movie content and extract similarities between different movies is of paramount importance for both movie service providers and their consumers [5]. Using movie recommendation systems allows us to identify our favorite movies in a reasonable amount of time. In addition, machine learning techniques are widely used today and play a vital role in many AI systems [6]. Therefore, we provide a movie recommendation system tailored to the audience's age group benefiting machine learning techniques that use information retrieval methods to make suggestions based on the hidden information contained in the products. Many systems provide such services, which can be classified into the following three categories:

- Collaborative filtering systems, e.g., MovieLens2, allow users to view similar movies and offer recommendations based on the tastes and preferences of other users who have viewed the existing films [7].
- In content-based systems, a movie is displayed with a set of features, which are mainly based on metadata such as the director, actor, genre, etc. The similarity between the movies is obtained from these features [8]. Jinni is one of the most complex systems in this group.
- Hybrid systems such as IMDB combine the previous two groups.

In order to create relevant representations of movies and to evaluate their similarity, these systems rely on man-made information about films. Content-based methods rely on movie metadata and build a database of this information for classification purposes. In other words, these systems do not take the raw content of the movie itself into account but are based solely on user-generated annotations [5].

In this research, we leverage subtitles to induce a new representation that encompasses latent conceptual

http://dx.doi.org/10.22133/ijwr.2022.370251.1139

Citation M. Kordabadi, A. Nazari, and M. Mansoorizadeh, " A Movie Recommender System based on Topic Modeling using Machine Learning Methods," *International Journal of Web Research*, vol.5, no.2,pp.19 -28, 2022, doi: 10.22133/ijwr.2022.370251.1139.

*Coressponding Author
Article History: Received: 14 November 2022; Revised: 25 December 2022 ; Accepted: 30 December 2022

information in the movies. A subtitle is a short piece of natural language text that can be exploited to estimate similarities between movies, much like to generic text documents. We are investigating the correlation between low-level textual similarities extracted from topic modeling and high-level relationships between the movies [9]. A movie can be recommended to users based on the user's age and the movie's topic simultaneously by extraction and clustering of these similarities. Available recommender systems usually overlook this aspect of the available information.

We collected a dataset consisting of 242 movies with different age groups from Subtitle Provider websites[1] . A set of topics is derived after text preprocessing and applying LDA topic modeling. By analyzing the topics, we can identify different aspects of their similarities. The proposed method considers not only the user's interests in the topics but also their age group which has been ignored in most of the existing systems. The results of the evaluation show that the use of information retrieval methods for this purpose is promising. The main contributions to this article include the following:

- Exploration of the movie information using LDA topic modeling
- Clustering documents and identifying their content similarity and predicting movie age ratings
- Development of a recommendation system based on the content of the film and its age group

The rest of this article is organized as follows. In Section 2 we will review the research literature, and in Section 3 the overall workflow, details of the proposed method and complementary techniques, are explained. Section 4 describes how to collect data and evaluate results. Section 5 concludes and outlines directions for further research.

2. A Review of Research Literature

Recently, many attempts have been made to adopt text documents in recommendation tasks [10]. In general, movie recommender systems can be divided into two categories: collaborative filtering systems and content-based systems. Collaborative filtering systems extract similarities between users based on their individual attributes, their preferences, and their interactions with items. Then, the recommendations are provided using this information. These systems can be divided into memory-based and model-based algorithms. Memory-based methods do not have a classic training phase. Much like lazy learning scenarios, they just memorize given users and measure the similarity of a new user with them by means of a lookup procedure such as neighborhood-based weighted averaging. Finally, recommendations are made using this summarized similarity. Model-based methods, on the other hand, aim to predict the user rating of a film by using predictive models. Popular approaches in this category are systems used by large companies such as Amazon, Netflix, etc. Amazon uses collaborative filtering systems to suggest products to its customers for at least a decade [3]. However, collaborative filtering systems generally entail high computational costs and perform poorly in the case of diverse data. Also, if users with similar tastes will score similarly is not necessarily precise. Content-based methods use content-related information and metadata to find similarities between them, without considering the behavior of the users. In content-based methods, textual, audio, and visual approaches are used to analyze and categorize movies. Some approaches, such as [11], use image and signal processing techniques to analyze audio and video features (video frames, audio clips, movie posters, etc.), while some textual properties (metadata such as plot summary, subtitles, genre, contributors, etc.) are analyzed through natural languages processing methods such as TF-IDF [12] and word2vec [13].

By analyzing the work performed in this area, it becomes evident that the text-based approach is less explored, which can be attributed to the lack of textual information regarding movies [14]. In addition, most textual information (such as transcripts, text, and oral captions) is mostly conversational. This implies that discovering textual information that reflects the topic and events of the movie is a challenging and significant task. The advantage of text-based methods is that human language contains more semantic information than visual or audio features, and text processing is computationally more comprehensive than image and audio processing [15]. There is also a myriad number of previous research on the textual classification of documents that can be used for text-based analysis. In fact, this strategy has several advantages over other approaches, despite the difficulty in obtaining useful textual content for movies. In addition, previous researchers have stated that textual information is more reliable and effective for video classification than video and audio [15].

Subtitles are the most common type of text (there are other texts to describe movies) for text-based classification of movies. The movie subtitle is a document with unstructured text. Obtaining the most common words is the easiest approach to retrieve information from documents. For example, the TF-IDF weights [16] can be used to indicate the relevance of each word in the text. It is possible to calculate content similarity by analyzing words extracted from a document instead of using numerical rankings in recommender systems [17]. Topics refer to important aspects of an item and the ability to extract topics from movie subtitles would assist a great deal in identifying content and discovering similarities between films. There are several methods for identifying topics in documents, including frequency-based, syntax-based, conditional random fields [18], and thematic modeling approaches such as LDA [19], and latent semantic analysis (LSA) [20]. Review topics can then be used to improve the actual ranking in recommender systems [21]. Further, they can be combined with latent factors in model-based CF [10] and with similarity criteria in neighbor-based CF [22]. The majority of research involving subtitles focuses on topic classification, especially news topics. In [23] and [24], subtitles are used to classify news items. The use of NLP techniques and the WordNet database allows topics to be categorized using subtitles [25]. Authors found that the accuracy of the classifier strongly depends on the intended topic tags. They are more accurate on topics such as sports, but less accurate on topics such as daily events because the categories are unclear or insufficient data is available. There are several studies that classify video topics using subtitles, but there are very few studies that use subtitles to categorize genres or classes. In [26], a new deep

[1] https://moviesubtitlesrt.com/; https://yts-subs.com/

neural architecture based on convolutional neural networks (ConvNets) was proposed to classify the film genres using a movie trailer. This method incorporates an incredibly deep ConvNet with the residual connections and uses a special convolution layer to extract time-based information from image-based features before mapping trailers into genres. In [27], the authors used film summaries to classify genres through the WordNet semantic similarity approach. In [28] IMDB-derived social tags are used to classify genres, with an accuracy rate of 50% to 65%. The authors employed information retrieval techniques to identify a particular film genre and then used the information gained in the previous step to recommend the film to users [29]. A content-based video recommendation system presented using the textual, audio, and video features of the film in [30]. Some studies, such as [31], applied film posters instead of text to identify the genre. In [32] Subtitles and introductory movies used to identify film genres (such as action, adventure, romance, etc.). In [26], using a deep neural network, film images examined to identify the genre of the film. In [33], the authors classified the movies using the text, sound, and image of the movies, along with a deep neural network, into two categories: suitable and unsuitable for children. In [34], a method for estimating the score (success) of the film presented based on a multi-dimensional latent variable model with poor supervision. [35] They benefited a block chain system to accurately rate movies based on user feedback.

In summary, the previous research is divided into three categories: 1- Identifying the topic similarity of the movies [12–15], [23, 36] 2- Predicting the film genre [24–29] and 3- Predicting the success rate (rating) [31–32]. However, no significant work has been done on age classification [18]. In this research, we present a method for age classification which examines the similarity of the content of movies to recommend them based on the topic and age category of the user.

3. The Proposed Method

An outline of the proposed method is presented in Figure 1. The proposed method includes the stages of text preprocessing, document vectorization, topic modeling of documents, clustering of documents, age classification, and movie recommendation.

3.1. Subtitle Preprocessing

In this step, preprocessing is performed on each text document and the unnecessary information is removed (information such as timestamps, markup elements, and many redundant and useless words). This process consists of the following steps:

1. Remove extra characters and symbols like timestamps, whitespace and punctuation marks.
2. Remove stop words such as “the”, “a”, “an”, “so”, “what”.
3. Remove named entities such as people, locations, and objects.
4. Tokenize the document into the list of its constituting terms.

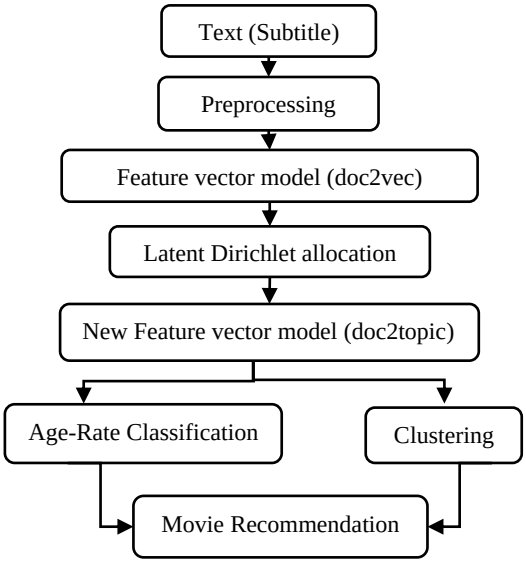


Figure. 1. Block diagram of the proposed method

Stemming and lemmatization: This step is a linguistic normalization technique and is used to convert words into their basic or root form.

3.2. Extracting Attributes (document-word)

At this stage, each document is represented as an object in a vector space. For this purpose, methods such as BoW, TFIDF, Word2Vec, and Bert can be used. In BoW, each document is represented as a multidimensional vector. This vector consists of the frequency words within the document. Collectively, the document vectors constitute the document-word matrix.

3.3. Topic Modeling

LDA and LSA are two common methods for topic modeling and discovering semantic relationships between documents. The LSA provides a simple way to extract the hidden semantic relations between documents. Nevertheless, it encounters many problems when dealing with words that have multiple meanings [9]. In order to resolve this problem, the topic modeling of LDA was proposed. The LDA is a topic model that assumes all documents are composed out of a set of latent topics. However, the distribution of words in each topic is different from others. With the help of the document-word matrix (BoW), the LDA can deduce these latent information (i.e., distribution of the topics in each document and the distribution of the words within each topic) [9]. We used the Gensim library for this purpose [37]. We implement two scenarios with 15 and 20 topics to evaluate the results.

3.4. Extract New Attributes (document-topic)

Due to the length of the vocabulary, the document-word matrix in the BoW is very large. This will increase the complexity of the model in machine learning algorithms. In order to address this problem, new features are extracted from the results of the LDA algorithm.

The output of the LDA algorithm is a matrix, T_{MxN}, where M is the number of documents and N is the number of topics (which is much smaller than the number of words). T_{ij} is the contribution of topic j in document i. Since machine learning

methods suffer from poor performance when dealing with large amounts of data, we use T instead of document-word matrix.

3.5. Clustering Documents

In this step the K-means algorithm is used to cluster the topics matrix, T. This algorithm operates based on the distance between points in space and places the documents inside a cluster that has the shortest distance from the center of that cluster.

3.6. Age Classification

The KNN, support vector machine, neural networks, logistic regression, multinomial Bayes, ID3 decision tree, AdaBoost, random forest, CNN and LSTM were used to classify the films according to the user classification. In this study, unlike other works in which the classification is done into two classes (suitable and unsuitable for children) [18], the movies are grouped into five known classes: General Audiences (G), Parental Guidance Suggested (PG), Parents Strongly Cautioned (PG-13), Restricted(R), and Adults Only (NC-17). The results of this step are used to build a list of recommended movies.

The age classification of films is based on the Film Censorship Act, which was approved by the United States Congress in 1930. Many countries adhere to the aforementioned law and broadcast the film in accordance with it. According to the same law, film production companies monitor the classification of produced films [38].

Movies are classified into five categories under broadcasting law.

1. Group G: A group of films that can be seen by all ages, which means that the film has no risk even for minors and children.
2. PG: It means the decision of the parents, maybe it is not suitable for the children, it means that the parents may not like this series to be seen by small children, and maybe they do not have a problem.
3. PG-13 group: The new PG-13 category has separated a group of movies from the PG group. The group does not allow children under the age of 13 to go to the cinema without their parents. In other words, PG-13 means a strong warning to parents. It means that this category of movies is not suitable for children under 13, and parents should be careful. It does not include group violence, sex, or nudity; the only common scenes are love and drugs.
4. NC_17 category: People under 17 years old are not allowed to watch this movie category. These videos may contain explicit sex, foul language, excessive violence, or all of these things. NC-17 does not mean the movie is gross or immoral.
5. Group R: These movies are suitable for over-17-year-olds; the R rating contains vulgar, violent, sexual, and drug use.

An average of 48 films were selected from each category.

The subtitles were in English and were downloaded from subtitle provider sites and manually checked for errors. After extracting the subtitle texts of the movies, we have the documents on which we perform information retrieval and thematic modeling of the movies.

3.7. Recommender System

After clustering movies and age categorizing, we recommend videos using the developed content-based recommendation system. The document-document similarity matrix further improves the results [39]. With this approach, movies that are appropriate for a user's age group and have similar content are recommended.

4. Evaluating the Results

In the proposed approach, we collected a real data set of 242 movies to demonstrate the utility of topic modeling in age rating and movie recommendation. These movies were selected from the movies on the popular website IMDb. The dataset is populated with different types of movies to avoid specific metadata biases, such as genre or actor (preventing data bias in a particular genre or category).

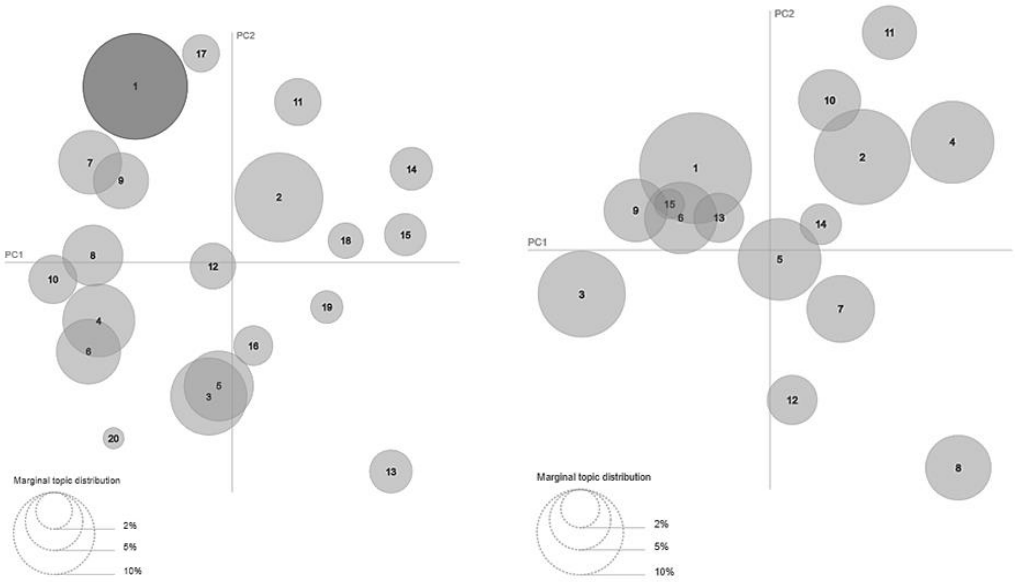
4.1. Clustering movies based on content

After identifying the topics in all documents, the documents were grouped according to the similarity of the topics. Figure 2 shows the distribution of 20 and 15 topics. In general, topics with fewer overlaps are more appropriate. In the case of 20 topics, a few of them overlap such that they can be merged to reduce complexity and computation. Table 1 also shows top contributing words of the topics. A document may include several topics, of which a few cases are dominant with significant topical densities.

Table 2 reports three major topics of a set of movies. For instance, topics 4, 17, and 8 constitute 40%, 27%, and 14% of “About Alex”. A document may include several topics. In the document-topic matrix each document is represented by 20 topics, but usually a few cases are dominant with significant topical densities. Out of 20 topics, we have selected 3 topics that each document is more relevant to as the dominant topics. Table 3 reports the dominant topics of some of the films. For instance, topics 4, 17, and 8 constitute 40%, 27%, and 14% of “About Alex”. “About Alex” is in category R, and topics 4 and 8 also contain words appropriate to this category. There are a number of dominant topics in more than one document, for example topic1 is in G and P G groups and topic8 is in R movies.

In Figure 3 each cluster is depicted by the frequency of members in each age group. The clusters suggest that movies with close age groups co-occur significantly. Experiments reveal that the purity of clusters increases with increasing the number of clusters. This clustering approach can be used to recommend movies to users based on both their age category and subject, simultaneously.

Clusters containing G and PG movies may rarely include movies from NC-17 and R groups and vice versa. However, due to the intrinsic similarities of groups PG-13, PG and NC-17, their cooccurrence is quite similar in the induced clusters. Table 3 shows representative movies for the clusters along with their major topics. A straight conclusion is that movies within a cluster share one or more major topics. For instance, Valiant and Charlottes from group G and Action Replay from



a: The distance between the topics (topic number=20) b: The distance between the topics (topic number=15)

Figure. 2. Distribution of topics

Table 1. Main words in each topic and their contribution

topic	*word1*	*word2*	*word3*	*word4*	*word5*
1	good	love	girl	school	talk
	0.21	0.18	0.15	0.11	0.08
2	music	dog	continue	instrumental	play
	0.14	0.13	0.11	0.09	0.09
3	move	man	kill	fire	hold
	0.18	0.15	0.11	0.10	0.07
4	people	film	sex	time	make
	0.21	0.17	0.14	0.12	0.08
5	leave	things	time	mom	give
	0.2	0.19	0.11	0.10	0.09
6	time	planet	fight	make	find
	0.16	0.14	0.11	0.11	0.10
7	time	good	play	space	money
	0.16	0.13	0.12	0.11	0.09
8	fuck	shit	man	guy	gone
	0.21	0.17	0.13	0.11	0.07
9	car	race	good	drive	start
	0.20	0.14	0.11	0.07	0.07
10	wait	boy	big	back	stop
	0.18	0.17	0.16	0.14	0.12
11	dream	back	guy	game	work
	0.15	0.13	0.13	0.13	0.11
12	kill	man	give	die	back
	0.21	0.16	0.14	0.12	0.09
13	man	baby	marriage	penny	girl
	0.21	0.18	0.13	0.10	0.06
14	laugh	scream	grunt	man	gasp
	0.17	0.16	0.11	0.10	0.08
15	dear	mother	family	pooh	wood
	0.17	0.16	0.16	0.13	0.12
16	good	guy	great	work	kid
	0.18	0.17	0.11	0.09	0.08
17	people	day	water	world	city
	0.19	0.18	0.17	0.15	0.14
18	love	night	back	feel	dance
	0.22	0.14	0.13	0.11	0.07
19	fly	eat	egg	bird	wait
	0.19	0.17	0.15	0.15	0.14
20	year	find	paint	head	master
	0.15	0.12	0.11	0.11	0.9

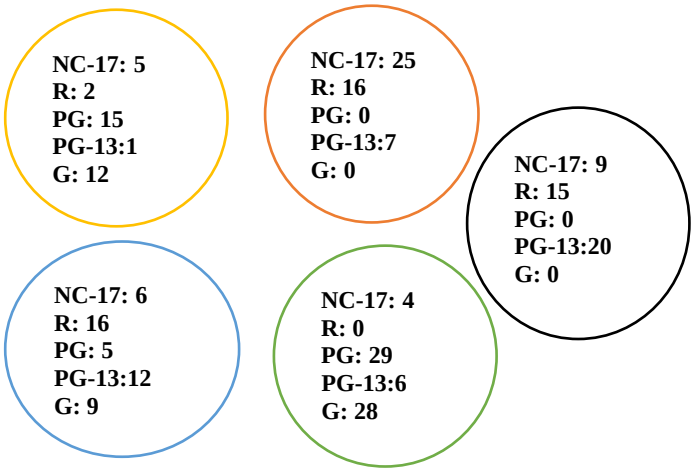


Figure. 3. Content clustering of movies

group PG share two major topics and all belong to cluster 4. Similarly, About Alex, Dutch Wife in the Desert, and Alex Cross share the major topic 17 and all belong to cluster 2.

4.2. Age Rating Classification

The KNN, SVM, MLP networks, logistic regression, Multinomial Bayes, ID3 decision tree, AdaBoost, Random Forest, CNN, and LSTM models were used. For the KNN model, 5 nearest neighbors are selected. The selection is based on an exhaustive inspection over the reasonable values of K. Furthermore, the adopted SVM model was equipped with Gaussian kernel to support nonlinear class boundaries.

For the MLP neural network model, we used three hidden layers with 150, 100, and 50 neurons. The Relu activation function is used for input and hidden layers. The softmax activation function is used for the output layer. RMSprop with an initial learning rate of 0.01 and a reduction of 0.00001 was used to optimize the model. Due to the nature of the problem, which is a multiclass categorization, the categorical cross entropy error function has been used.

In the employed CNN model, four fully connected layers, each with 100 neurons follow the basic convolutional modules. The dropout technique with ratio of 0.2 is used to regulate the network and reduce the generalization error. The optimizer function was Adam with a learning rate of 0.01.

LSTM, or Long-Short Term Memory, is a special type of recurrent neural network (RNN). In this network, there are three gates with a memory cell in each layer. Through these gates, the network controls the data flow and transfers features from the input to output layers.

Table 2. Dominant topic

id	*name*	*dominant topic 1*		*dominant topic 2*		*dominant topic 3*	
10	Dutch Wife in the Desert	7	0.18	15	0.14	4	0.13
50	About Alex	4	0.4	17	0.27	8	0.14
100	Action Replay	15	0.31	1	0.28	6	0.13
150	Alex Cross	17	0.22	12	0.16	15	0.11
200	Charlottes	3	0.35	1	0.12	6	0.11
240	Valiant	9	0.21	1	0.11	6	0.1

Table 3. Five representative movies for the clusters

Cluster#	*Movie*	*Age Rate*	*dominant topics*
1	Descent	NC-17	13,17,2
	Into the Woods	PG	6,17,2
	The Muppets	PG	20,1,6
	March of the Penguins	G	9,16,1
	Marche de l emperors La	G	9,16,1
2	Dutch Wife in the Desert	NC-17	7,15,4
	In the Realm of the Senses	NC-17	17,6,15
	About Alex	R	4,17,8
	Alex Cross	PG-13	17,12,15
	Camp Hope	R	15,17,16
3	Intent to Kill	NC-17	17,14,5
	Are You Scared	R	17,12,0
	Bad Ass	R	12,6,17
	Armored	PG-13	5,12,14
	Big Eyes	PG-13	10,14,17
4	Action Replay	PG	15,1,6
	Tangled	PG	1,6,2
	Charlottes	G	3,1,6
	Valiant	G	9,1,6
	Born to Race	PG-13	9,6,20
5	EN Ori	NC-17	9,12,5
	A Mighty Heart	R	5,4,8
	Aliens vs Predator Requiem	R	8,5,1
	Aeon Flux	PG-13	4,18,12
	Angels And Demons	PG-13	18,1,5

For evaluation of the results, we used accuracy, precision, and recall as the main criteria in the field of information retrieval. These measures determine the appropriateness of documents retrieved by the system. Tables 4 and 5 show the results of applying different algorithms with 20 and 15 topics, respectively.

After examining the results obtained by the implementation of the algorithms, it is evident that selecting 20 topics gives better results than selecting 15 topics. KNN, MLP, and Random Forest outperform the other models in all three evaluation criteria. Both CNN and LSTM networks outperform other approaches significantly in terms of accuracy, but they are not effective in terms of recall. Since the recall shows the ratio of related documents retrieved to the total number of related documents, in favor of higher coverage, it has been more important in the evaluation of IR systems.

We decided to use the one-against-all (OAA) classification scheme, which is an extension of two-class classification algorithms for multiple classes. This implementation involves dividing the multiclass data set into multiple binary classification problems. A binary classifier is then trained on each binary problem, and predictions are made using the model that is the most reliable. We have used the MLP neural network after examining several models. Tables 6 and 7, report the results obtained from different algorithms with 20 and 15 topics using the OAA method.

Comparing the results of the OAA method to other methods, it can be observed that the MLP and Random Forest methods perform better. Figure 4 shows another display of the results.

After clustering the content of the movies and the results of the age classification of the movies, it is possible to make suitable recommendations for the user by finding their age

Table 4. Age rating classification results (20 topic)

Classifier	*Precision*	*Recall*	*Accuracy*
KNN	0.68	0.61	0.63
MLP	0.56	0.53	0.53
SVM	0.58	0.49	0.48
LR	0.49	0.53	0.51
MultinomialNB	0.55	0.46	0.47
Decision Tree	0.5	0.51	0.49
AdaBoost	0.46	0.41	0.43
Random Forest	0.55	0.53	0.53
LSTM	0.43	0.39	0.85
CNN	0.63	0.37	0.87

Table 5. Age rating classification results (15 topic)

Classifier	*Precision*	*Recall*	*Accuracy*
KNN	0.56	0.55	0.55
MLP	0.41	0.41	0.40
SVM	0.41	0.41	0.40
LR	0.41	0.41	0.40
MultinomialNB	0.47	0.45	0.44
Decision Tree	0.44	0.43	0.43
AdaBoost	0.40	0.39	0.39
Random Forest	0.49	0.49	0.49
LSTM	0.36	0.26	0.82
CNN	0.58	0.34	0.84

Table 6. OAA results (20 topic)

Method	*Class*	*G*	*Pg-13*	*Pg*	*Nc-17*	*R*	*Average*
MLP	precision	0.95	0.94	0.92	0.93	0.93	0.934
	recall	0.95	0.94	0.92	0.92	0.92	0.93
	accuracy	0.95	0.94	0.92	0.92	0.92	0.93
LR	precision	0.87	0.87	0.85	0.85	0.84	0.856
	recall	0.85	0.85	0.84	0.85	0.82	0.842
	accuracy	0.85	0.85	0.84	0.85	0.82	0.842
KNN	precision	0.81	0.87	0.85	0.92	0.81	0.852
	recall	0.76	0.85	0.81	0.91	0.78	0.822
	accuracy	0.76	0.85	0.81	0.91	0.78	0.822
SVM	precision	0.84	0.68	0.79	0.82	0.73	0.772
	recall	0.84	0.67	0.78	0.82	0.71	0.764
	accuracy	0.84	0.67	0.78	0.82	0.71	0.764
NB	precision	0.92	0.8	0.83	0.83	0.8	0.836
	recall	0.91	0.79	0.8	0.83	0.73	0.812
	accuracy	0.91	0.79	0.77	0.83	0.7	0.8
ID3	precision	0.97	0.87	0.84	0.84	0.8	0.864
	recall	0.97	0.87	0.84	0.84	0.78	0.86
	accuracy	0.97	0.87	0.84	0.84	0.78	0.86
AdaBoost	precision	0.94	0.91	0.92	0.87	0.82	0.892
	recall	0.94	0.91	0.91	0.87	0.81	0.888
	accuracy	0.94	0.91	0.91	0.87	0.81	0.888
RF	precision	0.97	0.96	0.91	0.91	0.93	0.936
	recall	0.97	0.96	0.9	0.91	0.91	0.93
	accuracy	0.97	0.96	0.9	0.91	0.91	0.93

Table 7. OAA results (15 topic)

Method	*Class*	*G*	*Pg-13*	*Pg*	*Nc-17*	*R*	*Average*
MLP	precision	0.94	0.93	0.92	0.91	0.94	0.928
	recall	0.94	0.93	0.91	0.91	0.93	0.924
	accuracy	0.94	0.93	0.91	0.91	0.93	0.924
LR	precision	0.88	0.79	0.78	0.82	0.78	0.81
	recall	0.88	0.78	0.78	0.82	0.76	0.804
	accuracy	0.88	0.78	0.78	0.82	0.76	0.804
KNN	precision	0.93	0.86	0.84	0.8	0.83	0.852
	recall	0.91	0.85	0.82	0.8	0.78	0.832
	accuracy	0.91	0.85	0.82	0.8	0.78	0.832
SVM	precision	0.84	0.62	0.73	0.8	0.74	0.746
	recall	0.84	0.62	0.74	0.8	0.7	0.74
	accuracy	0.84	0.62	0.74	0.8	0.7	0.74
NB	precision	0.89	0.78	0.86	0.86	0.86	0.85
	recall	0.89	0.77	0.85	0.85	0.85	0.842
	accuracy	0.89	0.76	0.83	0.83	0.83	0.828
ID3	precision	0.87	0.81	0.89	0.9	0.8	0.854
	recall	0.87	0.81	0.89	0.9	0.8	0.854
	accuracy	0.87	0.81	0.89	0.9	0.8	0.854
AdaBoost	precision	0.87	0.81	0.86	0.88	0.84	0.852
	recall	0.87	0.81	0.85	0.88	0.79	0.84
	accuracy	0.87	0.81	0.86	0.88	0.79	0.842
RF	precision	0.94	0.92	0.92	0.92	0.91	0.922
	recall	0.94	0.92	0.92	0.91	0.91	0.92
	accuracy	0.94	0.92	0.92	0.91	0.91	0.92

and taking their age category as well as their favorite movies into account.

The rest of this section compares the proposed method with several related works. In [40], recommendations for movies and restaurants are given based on each person's Google search history. The authors use LDA on Wikipedia corpus to extract topics, and then pair the topics with the subjects extracted from their browsing histories to extract user preferences. In [41], the authors use Naive Bayes algorithm to investigate user reviews and extract their sentiments, in order to predict popularity of the movies. We adopt this work in our comparisons; since sentiment analysis-based movie labeling can be benefited as a base for movie recommendations. Authors in [42] used script text to analyze violent content in movies. Their approach is based on a wide range of features designed to capture lexical, semantic, sentiment, and offensive language aspects. We selected this work for comparison because it tries to categorize user ages, which is correlated to the level violent content. The authors of [33] suggest an RNN-based architecture for predicting age ratings, paying special attention to the joint modeling of genre and emotions in the script. Table 8 demonstrates our approach compared to the mentioned works.

5. Discussion

The used learning models have been successfully applied in information retrieval problems, especially text classification. These models have very good accuracy for two-class tasks that can be separated by linear classifiers. By examining the results obtained from the implementation of the mentioned algorithms, it is evident that when we extract 20 topics, better results have been obtained than when we choose 15 topics, hence it can be concluded that choosing the right number of topics in the semantic modeling stage has improved retrieval results. Among the models, KNN, MLP, and random forest show better results in all three evaluation criteria. CNN and LSTM networks show significantly higher accuracies than other methods, however in terms of recall, they do not perform well. In information retrieval systems, usually the recall is treated more importantly, since it extracts as much information as possible. In the one-versus-all (one-versus-rest) multiclass classification method, the importance of the appropriate number of topics can be inspected by comparing results of 20 and 15 topics; where again superior results come from 20 topics. Another point worth mentioning is that random forest and MLP methods result in better precision, recall, and accuracy in both of the experiments. Furthermore, data augmentation has been very effective in achieving better results.

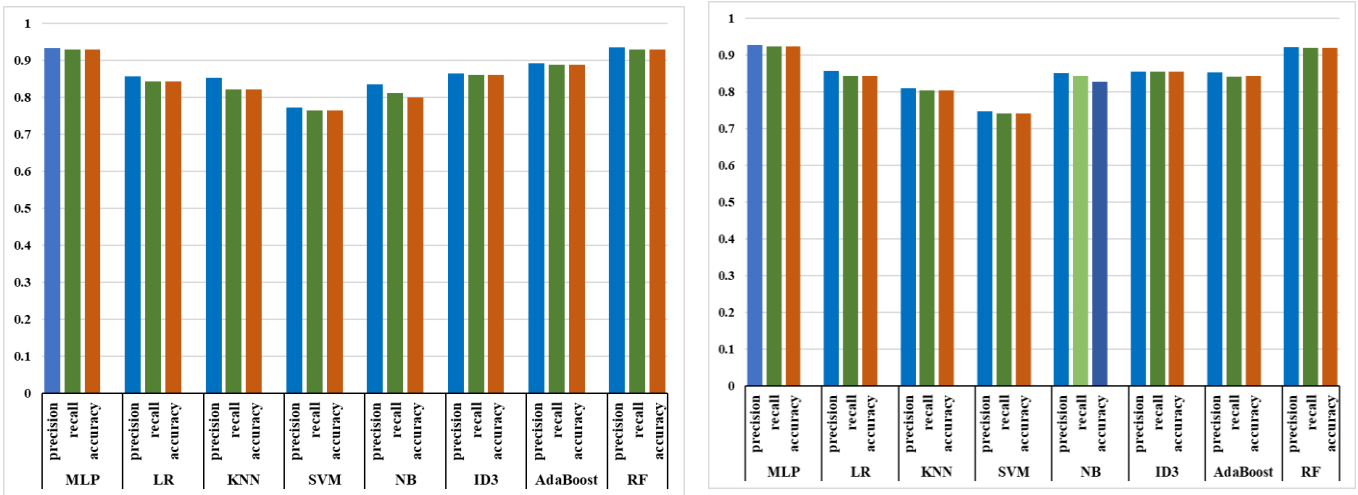
6. Conclusion

In this paper, information retrieval and topic modeling methods were used to predict the classification and clustering of movies based on content and topic to suggest the movies. Additionally, the proposed method pays attention to the age group of the user, as well as the user's interest in the topic. This issue has been overlooked in most previous research and recommender systems. The results of the evaluation show that information retrieval methods can be used to achieve this goal. Selecting the appropriate number of topics in the semantic modeling stage enhances the results in information retrieval and the recommendations. The proposed method, in addition to the appropriate movie suggestion, can be used to identify the exact age category of the movies.

For future work, we intend to apply the proposed approach to the neighboring questions in movie industry such as predicting the success or failure of movie sales based solely on the script, in the early pre-production stages. Since the script is available in advance, we can estimate the age group, similarity between available films, and the target audience of the film in early production stages.

Table (8): Comparison of different methods

Ref	*Precision*	*Recall*	*Accuracy*
[40]	51%	58%	-
[41]	79.65%	79.65%	79.44%
[42]	60.9%	60.0%	-
[33]	93%	87%	88%
Proposed	93.4%	93%	93%



a: Topics=20 b: Topics=15

Figure. 4. Average of results

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

MK: Study design, acquisition of data, interpretation of the results, statistical analysis, drafting the manuscript. AM: Study design, acquisition of data, interpretation of the results, statistical analysis, drafting the manuscript. MM: Supervision, interpretation of the results, revision of the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

References

[1] Shafaei, M., et al., Rating for parents: Predicting children suitability rating for movies based on language of the movies. arXiv preprint arXiv:1908.07819, 2019.

[2] Hofstätter, S., et al., Improving efficient neural ranking models with cross-architecture knowledge distillation. arXiv preprint arXiv:2010.02666, 2020.

[3] Chen, S., et al. Fine-grained video-text retrieval with hierarchical graph reasoning. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[4] Singla, R., et al. FLEX: a content based movie recommender. in 2020 International Conference for Emerging Technology (INCET). 2020. IEEE.

[5] Goyani, M. and N. Chaurasiya, A Review of Movie Recommendation System. ELCVIA: electronic letters on computer vision and image analysis, 2020. 19(3): p. 18-37.

[6] Katarya, R. and O.P. Verma, An effective collaborative movie recommender system with cuckoo search. Egyptian Informatics Journal, 2017. 18(2): p. 105-112.

[7] Subramaniyaswamy, V., et al., A personalised movie recommendation system based on collaborative filtering. International Journal of High Performance Computing and Networking, 2017. 10(1-2): p. 54-63.

[8] Reddy, S., et al., Content-based movie recommendation system using genre correlation, in Smart Intelligent Computing and Applications. 2019, Springer. p. 391-397.

[9] Jelodar, H., et al., Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 2019. 78(11): p. 15169-15211.

[10] Srifi, M., et al., Recommender systems based on collaborative filtering using review texts—A survey. Information, 2020. 11(6): p. 317.

[11] Lehinevych, T., et al. Discovering similarities for content-based recommendation and browsing in multimedia collections. in 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems. 2014. IEEE.

[12] Cataltepe, Z., M. ULUYAĞMUR, and E. TAYFUR, Feature selection for movie recommendation. Turkish Journal of Electrical Engineering & Computer Sciences, 2016. 24(3): p. 833-848.

[13] Chen, H.-W., et al. Fully content-based movie recommender system with feature extraction using neural network. in 2017 International conference on machine learning and cybernetics (ICMLC). 2017. IEEE.

[14] Ibrahim, Z.A.A., S. Haidar, and I. Sbeity, Large-scale Text-based Video Classification using Contextual Features. European Journal of Electrical Engineering and Computer Science, 2019. 3(2).

[15] Khattar, D., et al. Mvae: Multimodal variational autoencoder for fake news detection. in The world wide web conference. 2019.

[16] Salton, G. and C. Buckley, Term-weighting approaches in automatic text retrieval. Information processing & management, 1988. 24(5): p. 513-523.

[17] Terzi, M., et al. Text-based user-knn: Measuring user similarity based on text reviews. in International Conference on User Modeling, Adaptation, and Personalization. 2014. Springer.

[18] Xia, H., et al., Sentiment analysis for online reviews using conditional random fields and support vector machines. Electronic Commerce Research, 2020. 20(2): p. 343-360.

[19] Zoghbi, S., I. Vulić, and M.-F. Moens, Latent Dirichlet allocation for linking user-generated content and e-commerce data. Information Sciences, 2016. 367: p. 573-599.

[20] Chehal, D., P. Gupta, and P. Gulati, Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations. Journal of Ambient Intelligence and Humanized Computing, 2021. 12(5): p. 5055-5070.

[21] Qiu, L., et al., Aspect-based latent factor model by integrating ratings and reviews for recommender system. Knowledge-Based Systems, 2016. 110: p. 233-243.

[22] Wang, H. and N. Luo, Collaborative filtering enhanced by user free-text reviews topic modelling. 2014.

[23] Anwar, A., G.I. Salama, and M. Abdelhalim. Video classification and retrieval using arabic closed caption. in ICIT 2013 The 6th International Conference on Information Technology VIDEO. 2013.

[24] Lee, C.G., Text-based video genre classification using multiple feature categories and categorization methods. 2017.

[25] Katsiouli, P., V. Tsetsos, and S. Hadjiefthymiades. Semantic Video Classification Based on Subtitles and Domain Terminologies. in KAMC. 2007.

[26] Wehrmann, J. and R.C. Barros, Movie genre classification: A multi-label approach based on convolutions through time. Applied Soft Computing, 2017. 61: p. 973-982.

[27] Fourati, M., A. Jedidi, and F. Gargouri. Automatic identification Genre of audiovisual documents. in 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA). 2014. IEEE.

[28] 28. Hong, H.-Z. and J.-I.G. Hwang. Multimodal PLSA for movie genre classification. in International Workshop on Multiple Classifier Systems. 2015. Springer.

[29] Saumya, S., J. Kumar, and J.P. Singh, Genre fraction detection of a movie using text mining, in Advanced Computing and Systems for Security. 2018, Springer. p. 167-177.

[30] Bougiatiotis, K. and T. Giannakopoulos, Enhanced movie content similarity based on textual, auditory and visual information. Expert Systems with Applications, 2018. 96: p. 86-102.

[31] Kundalia, K., Y. Patel, and M. Shah, Multi-label movie genre detection from a movie poster using knowledge transfer learning. Augmented Human Research, 2020. 5(1): p. 1-9.

[32] Mangolin, R.B., et al., A multimodal approach for multi-label movie genre classification. Multimedia Tools and Applications, 2020: p. 1-26.

[33] Shafaei, M., et al., A Case Study of Deep Learning Based Multi-Modal Methods for Predicting the Age-Suitability Rating of Movie Trailers. arXiv preprint arXiv:2101.11704, 2021.

[34] Watanabe, K., et al. Movie Rating Estimation Based on Weakly Supervised Multi-modal Latent Variable Model. in 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE). 2021. IEEE.

[35] Maragatham, G. Movie Rating System based on Blockchain. in 2021 International Conference on Computer Communication and Informatics (ICCCI). 2021. IEEE.

[36] Luhmann, J., M. Burghardt, and J. Tiepmar, SubRosa: Determining Movie Similarities based on Subtitles. INFORMATIK 2020, 2021.

[37] Rehurek, R. and P. Sojka. Software framework for topic modelling with large corpora. in In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. 2010. Citeseer.

[38] Film Rating," Motion Picture Association of America, [Online]. Available: http://www.mpaa.org/ratings. [Accessed 05 08 2022].

[39] Chambua, J. and Z. Niu, Review text based rating prediction approaches: preference knowledge learning, representation and utilization. in Artificial Intelligence Review, 2021. 54(2): p. 1171-1200.

[40] Rajendran, D P D. and Rangaraja P. S. Using topic models with browsing history in hybrid collaborative filtering recommender system: Experiments with user ratings. in International Journal of Information Management Data Insights, 2021. 1(2): 100027.

[41] Samsir, S., et al. Implementation Naïve Bayes Classification for Sentiment Analysis on Internet Movie Database. in Building of Informatics, Technology and Science (BITS), 2022. 4 (1): p. 1-6.

[42] Martinez, Victor R., et al. Violence rating prediction from movie scripts. in Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

Mojtaba Kordabadi received his B.Sc. in software engineering in 2009 from Hamedan University of Applied Sciences. He graduated with a MSc degree in artificial intelligence from Buali Sina University, Hamedan. He is a teacher of computer courses at Hamedan Technical and Vocational University. His research interests include machine learning, recommender systems, data mining.

Amin Nazari received BSc degree in Computer Software Engineering from Islamic Azad University, Hamedan, in 2009. He received his MSc degree in Computer Software Engineering from Arak University, Arak, in 2015. He is now a Ph.D. candidate of artificial intelligence at the Bu-Ali Sina University, Hamedan. His research interests include wireless sensor networks, the Internet of Things, IoT-fog networks and recommender systems.

Muharram Mansoorizadeh is an associate professor at the Computer Engineering Department of Bu-Ali Sina University. He received his BSc degree in software engineering from the University of Isfahan, Isfahan, Iran, in 2001, and his MSc degree in software engineering and the PhD in computer engineering from Tarbiat Modares University, Tehran, Iran, in 2004 and 2010, respectively. His current research interests include machine learning, affective computing and information retrieval.