MOVIE REVIEWS

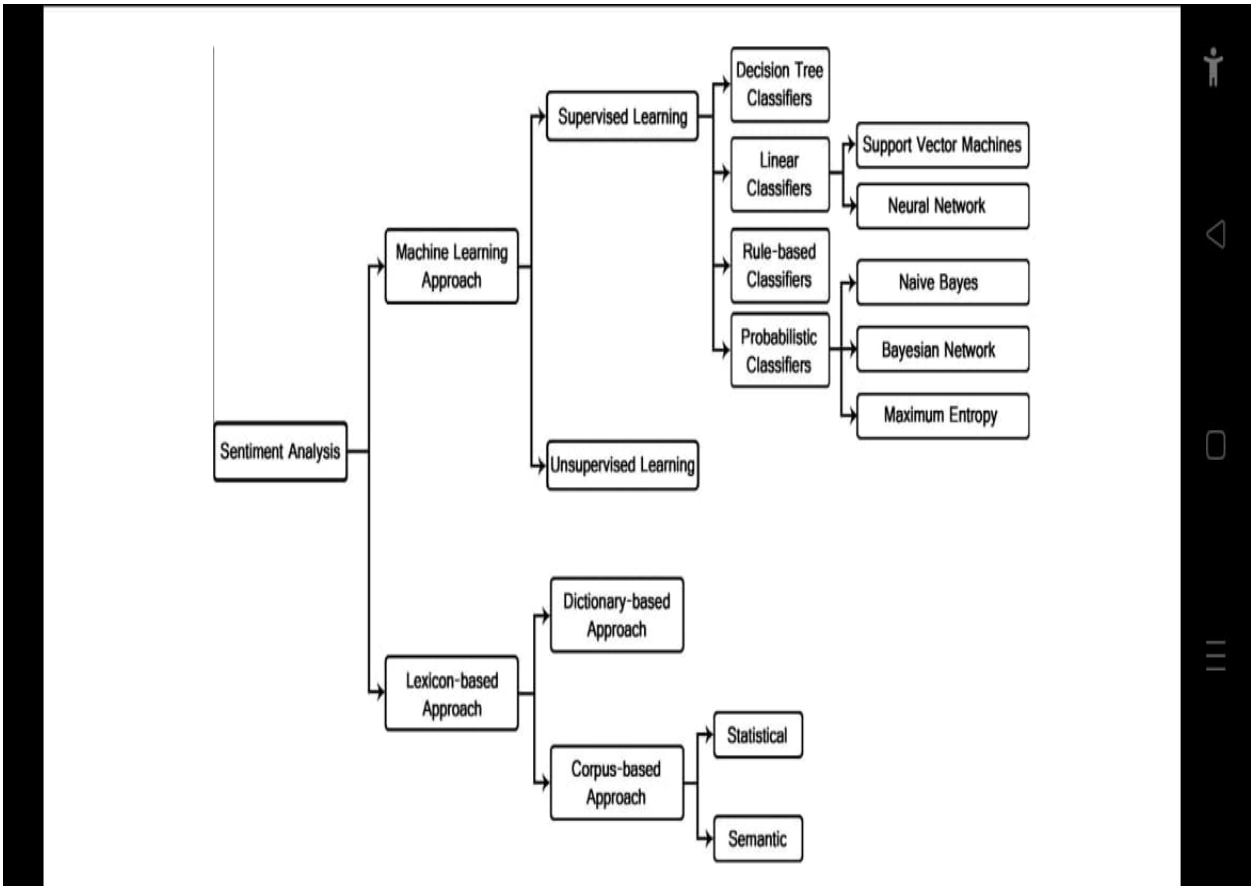
GP

NOVEMBER 6, 2022

Dr. Ahmed Elsayed

Abstract:

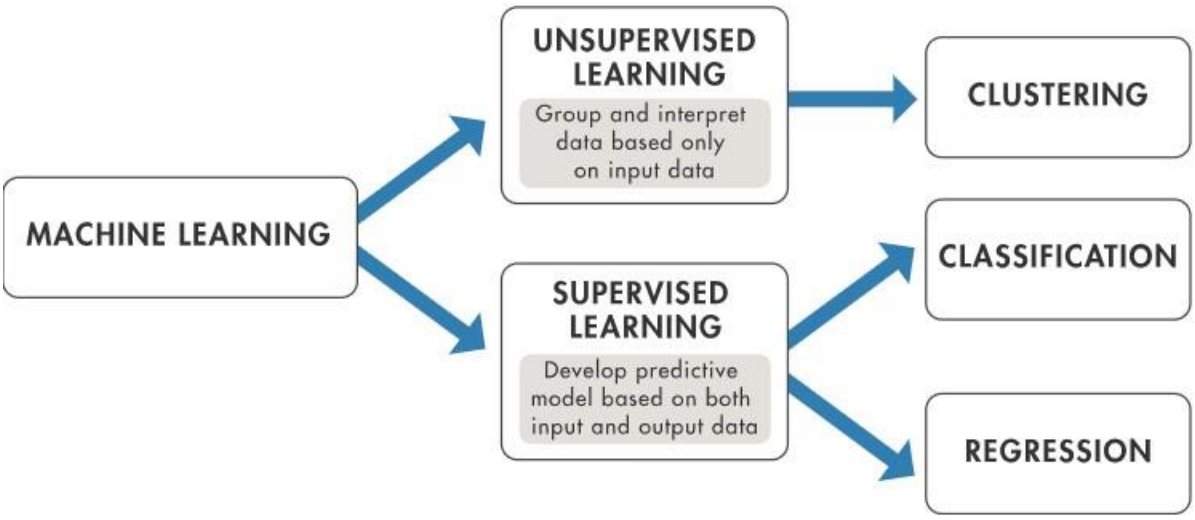
"MACHINE LEARNING APPROACH"



- **What Is Machine Learning?**

How it works, why it matters, and getting started
Machine Learning is an AI technique that teaches computers to learn from experience. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. Deep learning is a specialized form of machine learning.

How Machine Learning Works
Machine learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data



- **<u>Supervised Learning</u>**:

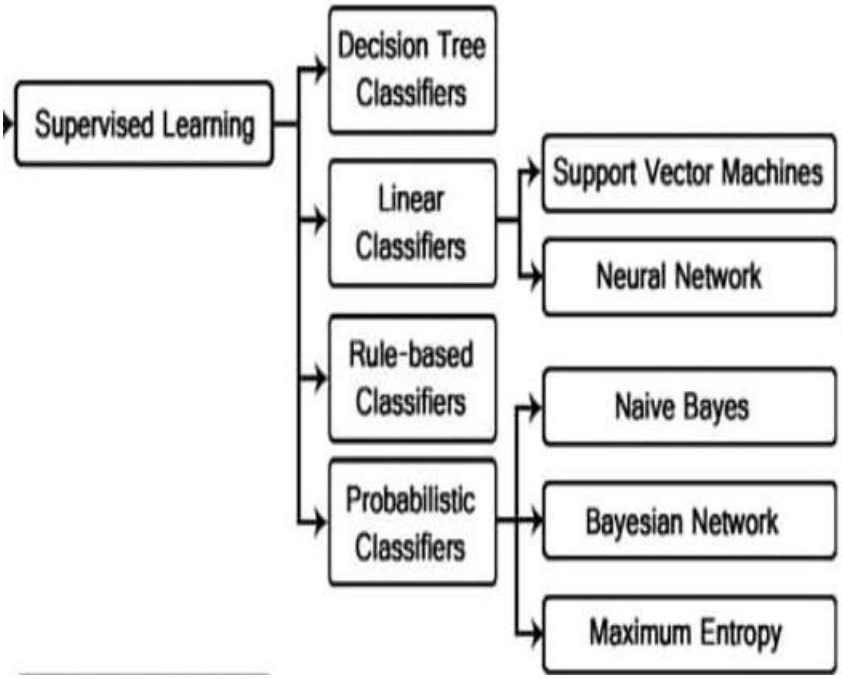
Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Use supervised learning if you have known data for the output you are trying to predict.

Supervised Learning: also known, Learn how supervised learning works and how it can be used to build highly accurate machine learning models.
What is supervised learning?
Supervised learning, as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.

How supervised learning works:

Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

Supervised learning can be separated into two types of problems when data mining—classification and regression:



Supervised learning uses classification and regression techniques to develop machine learning models.

Classification techniques:

predict discrete responses—for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring.
Use classification if your data can be tagged, categorized, or separated into specific groups or classes. For example, applications for hand-writing recognition use classification to recognize letters and numbers. In image processing and computer vision, unsupervised pattern recognition techniques are used for object detection and image segmentation. The most common algorithms for performing classification can be found here.

Regression techniques:

predict continuous responses—for example, hard-to-measure physical quantities such as battery state-of-charge, electricity load on the grid, or prices of financial assets. Typical applications include virtual sensing, electricity load forecasting, and algorithmic trading.
Use regression techniques if you are working with a data range or if the nature of your response is a real number, such as temperature or the time until failure for a piece of equipment. The most common algorithms for performing regression can be found

Naive Bayes:

Naive Bayes is classification approach that adopts the principle of class conditional independence from the Bayes Theorem. This means that the presence of one feature does not impact the presence of another in the probability of a given outcome, and each predictor has an equal effect on that result. There are three types of Naïve Bayes classifiers: Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes. This technique is primarily used in text classification, spam identification, and recommendation systems.

***Linear regression**:

Linear regression is used to identify the relationship between a dependent variable and one or more independent variables and is typically leveraged to make predictions about future outcomes. When there is only one independent variable and one dependent variable, it is known as simple linear regression. As the number of independent variables increases, it is referred to as multiple linear regression. For each type of linear regression, it seeks to plot a line of best fit, which is calculated through the method of least squares. However, unlike other regression models, this line is straight when plotted on a graph.

***Logistic regression**:
While linear regression is leveraged when dependent variables are continuous, logistical regression is selected when the dependent variable is categorical, meaning they have binary outputs, such as "true" and "false" or "yes" and "no." While both regression models seek to understand relationships between data inputs, logistic regression is mainly used to solve binary classification problems, such as spam identification.

***Support vector machine (SVM)**:

A support vector machine is a popular supervised learning model developed by Vladimir Vapnik, used for both data classification and regression. That said, it is typically leveraged for classification problems, constructing a hyperplane where the distance between two classes of data points is at its maximum. This hyperplane is known as the decision boundary, separating the classes of data points (e.g., oranges vs. apples) on either side of the plane.

***K-nearest neighbor:**

K-nearest neighbor, also known as the KNN algorithm, is a non-parametric algorithm that classifies data points based on their proximity and association to other available data. This algorithm assumes that similar data points can be found near each other. As a result, it seeks to calculate the distance between data points, usually through Euclidean distance, and then it assigns a category based on the most frequent category or average.

Its ease of use and low calculation time make it a preferred algorithm by data scientists, but as the test dataset grows, the processing time lengthens, making it less appealing for classification tasks. KNN is typically used for recommendation engines and image recognition.

Random forest:

Random forest is another flexible supervised machine learning algorithm used for both classification and regression purposes. The "forest" references a collection of uncorrelated decision trees, which are then merged together to reduce variance and create more accurate data predictions.

***Decision Tree and Boosted Tree (Gradient Boosting Machine)**

A decision tree uses a flowchart structure that typically contains a root, internal nodes, branches, and leaves. The internal node is where the attribute in question (eg, creatinine >1 or creatine <1) is tested, while the branch is where the outcome of this tested question is then delegated. The leaves are where the final class label is assigned which, in short, represents the final decision after it has incorporated the results of all the attributes.36, 37, 38, 39 The end result of the decision tree is a set of rules that governs the path from the root to the leaves. Simple decision trees are not commonly used in ML. However, variations such as the Gradient boosting machine is used for both classification and regression tasks.40, 41 Gradient boosting machine is an ensemble method that uses weak predictors (eg, decision trees) that can ultimately be boosted and lead to a better performing model (ie, the boosted tree). This method can sometimes yield ve11ry reasonable models, especially with unbalanced data sets. However, their limited number of tuning parameters may make them more prone to overfitting compared to RF that contains a larger number of parameters for tuning and finding the optimized model

Unsupervised learning:

Unsupervised learning refers to the use of artificial intelligence (AI) algorithms to identify patterns in data sets containing data points that are neither classified nor labeled.

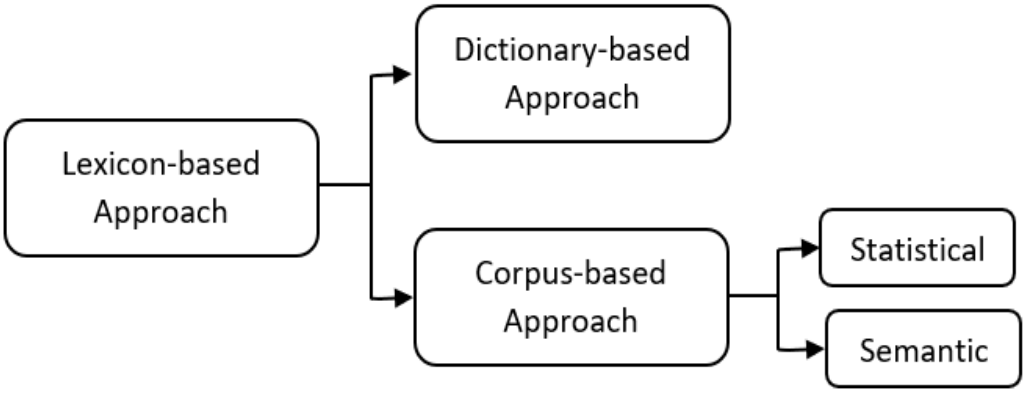
The algorithms are thus allowed to classify, label and/or group the data points contained within the data sets without having any external guidance in performing that task.

In other words, unsupervised learning allows the system to identify patterns within data sets on its own.
Lexicon Based Approach

Application of a lexicon is one of the two main approaches to sentiment analysis and it involves calculating the sentiment from the semantic orientation of word or phrases that occur in a text.

With this approach a dictionary of positive and negative words is required, with a positive or negative sentiment value assigned to each of the words. Different approaches to creating dictionaries have been proposed, including manual and automatic approaches.

Generally speaking, in lexicon-based approaches a piece of text message is represented as a bag of words. Following this representation of the message, sentiment values from the dictionary are assigned to all positive and negative words or phrases within the message. A combining function, such as sum or average, is applied in order to make the final prediction regarding the overall sentiment for the message. [1]



Lexicons

are dictionaries or vocabulary created specifically for use in sentiment analysis.[2]
This technique calculates the sentiment orientations of the whole document or set of sentences from semantic orientation of lexicons. Semantic orientation can be positive, negative, or neutral. The dictionary of lexicons can be created manually as well as automatically generated.[3]

in lexicon-based sentiment analysis, words in texts are labeled as
positive or negative (and sometimes as neutral) with the help of a so-called valence dictionary.

EX: Take the phrase “Good people sometimes have bad days.”. A valence dictionary would label the word “Good” as positive; the word “bad” as negative; and possibly the other words as neutral. Once each word in the text is labeled, we can derive an overall sentiment score by counting the numbers of positive and negative words, and combining these values mathematically.

If the sentiment score is negative, the text is classified as negative. It
follows that a positive score means a positive text, and a score of zero
means the text is classified as neutral.[8]
Apart from a sentiment value, the aspect of the local context of a word is usually taken into consideration, such as negation, intensity.

Dictionary-based

A sentiment analysis dictionary contains information about the emotions or polarity expressed by words, phrases, or concepts. In practice, a dictionary usually provides one or more scores for each word. We can then use them to compute the overall sentiment of an input sentence based on individual words.
In this approach, a dictionary is created by taking a few words initially. Then an online dictionary, thesaurus or WordNet can be used to expand that dictionary by incorporating synonyms and antonyms of those words. The dictionary is expanded till no new words can be added to that dictionary. The dictionary can be refined by manual inspection.[3]
Applications on Dictionary-based:

Amazon

Amazon is an American multinational technology company focusing on e-commerce, cloud computing, online advertising, digital streaming, and artificial intelligence. It has been referred to as "one of the most influential economic and cultural forces in the world, and is one of the world's most valuable brands. Amazon uses a rating system of 1 to 5 stars for the product. Where reviews are categorized into two groups: negative and positive, reviews of 1 star, 2 stars or 3 stars are negative, and 4 stars or 5 stars reviews are classified as positive. But this evaluation is inaccurate because it does not explain the consumer's opinion of the product accurately. The evaluation may be bad for a reason other than the product, such as the delay in delivering the product to the consumer. [2]

Table1: Algorithm results on Amazon products

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Confusion Matrix [[TP, FN], [FP, TN]]
Pattern Lexicon	69	88	72	79	[[25748,10204], [3350,4318]]
VADER Lexicon	83	90	89	89	[[31828,4124], [3421, 4247]]
SentiWordNet Lexicon	80	88	88	88	[[31653,4299], [4238, 3430]]

ENulog

eNulog is a collection of data for 6000 posts from 10 popular movie blogs. eNulog uses a lexical approach in which each word in tans is compared to the dictionary and is evaluated by rating

the word in the dictionary as positive or negative. If the sum of the points is positive then the text is positive and if it is negative then the text is negative. WordNet is used for these two lexical approaches and is the most famous dictionary and contains many solutions for lexical dictionaries with its accuracy reaching 82%.[4]

Twitter

Twitter is a microblogging and social networking service owned by American company Twitter, Inc., on which users post and interact with messages known as "tweets". Senticircles is a lexicon-based approach to Twitter sentiment analysis, it differs from typical lexical-based approaches. Senticircles finds common patterns of words in different contexts in tweets. Twitter datasets are evaluated using three different emotion dictionaries to derive emotion for the word. For sentiment detection at the tweet level, this method is 5-4% better than SentiStrength in terms of accuracy. [5]

IMDB

IMDB is an online database of information related to films, television series, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews. IMDB uses WordNet. Film reviews are divided into dictionaries for positive and negative words. Film reviews taken from IMDB contain 12,500 words. These words are compared to WordNet to see if the rating is positive or negative.
For the small size of the data set, the data set-based method is more useful in classifying opinions. Other datasets can use WordNet. [6]

Corpus-based

This finds sentiment orientation of context-specific words. The two methods of this approach are:

Statistical approach

The words which show erratic behavior in positive behavior are considered to have positive polarity. If they show negative recurrence in negative text they have negative polarity. If the frequency is equal in both positive and negative text then the word has neutral polarity. [3]

Semantic approach

This approach assigns sentiment values to words and the words which are semantically closer to those words; this can be done by finding synonyms and antonyms with respect to that word. [3]

Applications on Statistical approach:

Facebook

Facebook uses sentiment analysis of messages written by users to see if these comments are positive, negative or neutral. This method was implemented in sent Buk which is the Facebook app. sent Buk retrieves and categorizes messages written by users on Facebook. The results show that conducting sentiment analysis on Facebook has a high accuracy, representing 83%

Sentiment Analysis of Restaurant Customer

Reviews on TripAdvisor using (Naive Bayes,textblod)

Based on the result ofthis paper that done on restaurant customer reviews in Surabaya, customer satisfaction analysis might be used by the Naive Bayes classification method and TextBlob sentiment analysis that might be able to learn sentiment from customers, since customer satisfaction is essential in terms ofthe restaurant business. Accuracy Naive Bayes = 72.06% Accuracy TextBlob = 69.12%The results also indicate that the Naive Bayes method has a value of 72.06% accuracy and is slightly better (2.94%) than TextBlob sentiment analysis. Further research can be done by increasing the number and variety ofreview data, or by other methods, to increase the value of accuracy.

Restaurant reviews classification using NLP Techniques

In this paper, we proposed machine learning NLP techniques for classification of restaurant reviews .We removed stop words from the given dataset and apply stemming for efficiency. Afterthat , we applied three vectorization techniques count vectorizer ,TFIDF vectorizer ,Hashing vectorizer for converting textual data into numeric data .Later we applied machine learning models with these three vectorization mothods .We achieved an accuracy of 88% with TFIDF vectorization method and Logistic Regression classification model.

Reference

Links

1-
https://www.ibm.com/cloud/learn/supervised-learning
2-
https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
3-
https://www.techtarget.com/searchenterpriseai/definition/supervised-learning
4-
https://www.guru99.com/unsupervised-machine-learning.html

books

[1]Amazom Product Sentiment Analysis Using Machine Learning Techniques – Sobia Wassan & Noor Zaman – 2021 – pages 4,5,6.

[2] Sentiment Analysis on Consumer Reviews of Amazon Products – ching yuh uang – 2021 – pages 3.

[3]Social Network Analytics:Computational Research Methods and Techniques – Nilanjan Dey – 2019 – pages 143,144.

[4]Computers in Human Behavior – Professor Mattnieu Guitton – 2014 – pages 527,528,529,530

[5] security-informatics.springeropen.com - Anna Jurek , Maurice D. Mulvenna , Yaxin Bi – 2015.

[6] Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches - Heidi Nguyen, Aravind Veluchamy, Mamadou Diop, Rashed Iqbal - 2018 – pages 6-7,12,16-18.

[7]Application and techniques of opinion mining - Neha Gupta, Rashmi Agrawal – 2020 – pages 2-3.

[8]Advances in Artificial Intelligance - R. Goebel, J. Siekmann, W. Wahlster – 2008 – pages 37-38,43.

[9]Information Processing & Management – Jim Jansen – 2016 – pages 5-19.

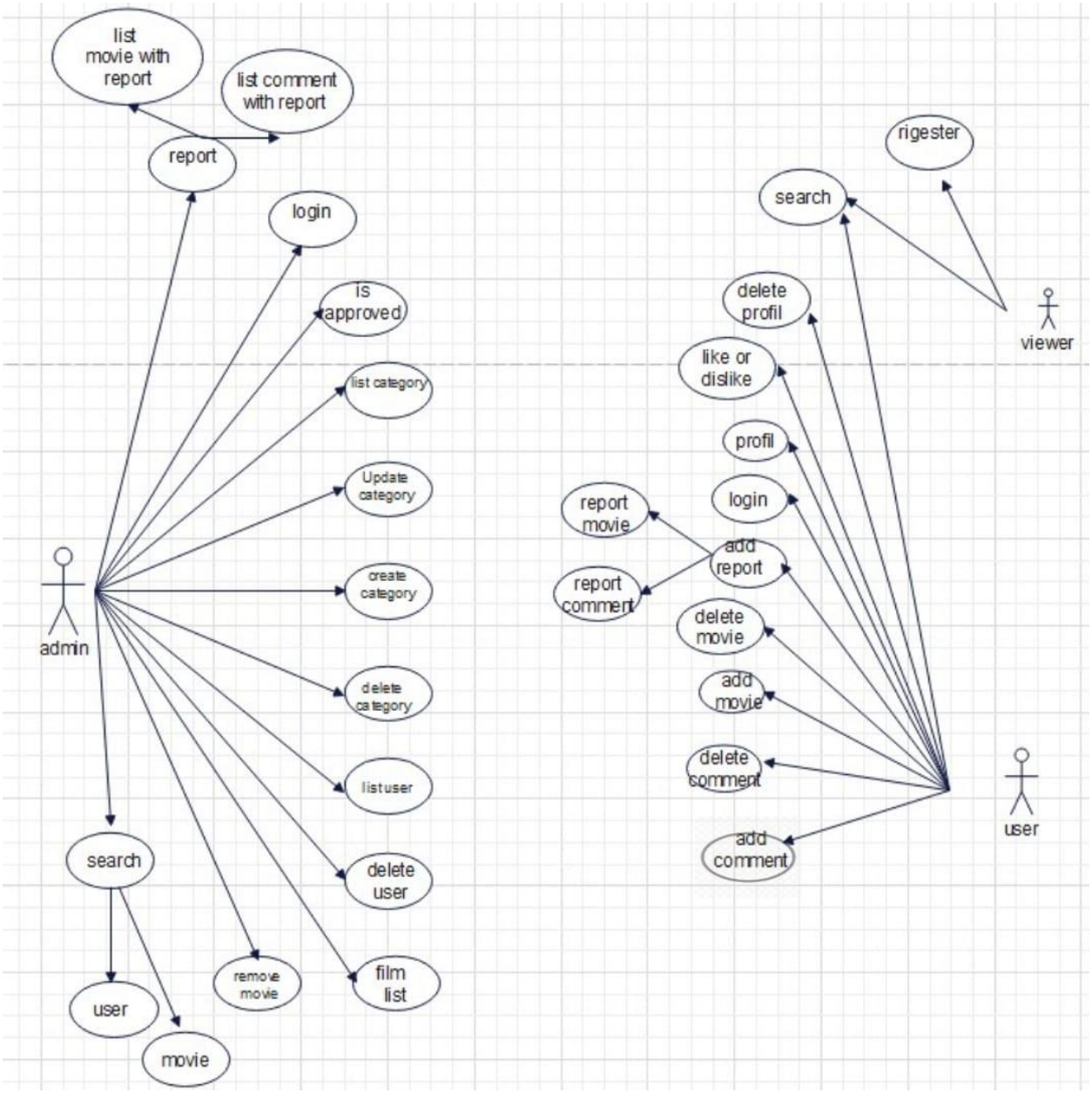
[10] Sentiment analysis on IMDB using lexicon and neural networks - Zeeshan Shaukat, Abdul Ahad Zulfqar, Chuangbai Xiao, Muhammad Azeem, Tariq Mahmood – 2020 – pages 4-5.

[11]Computers in Human Behavior – Professor Mattnieu Guitton – 2014 – pages 527-530.

[12] Knime.com - Aline Bessa – 2022.

Chapter 3: Software Design

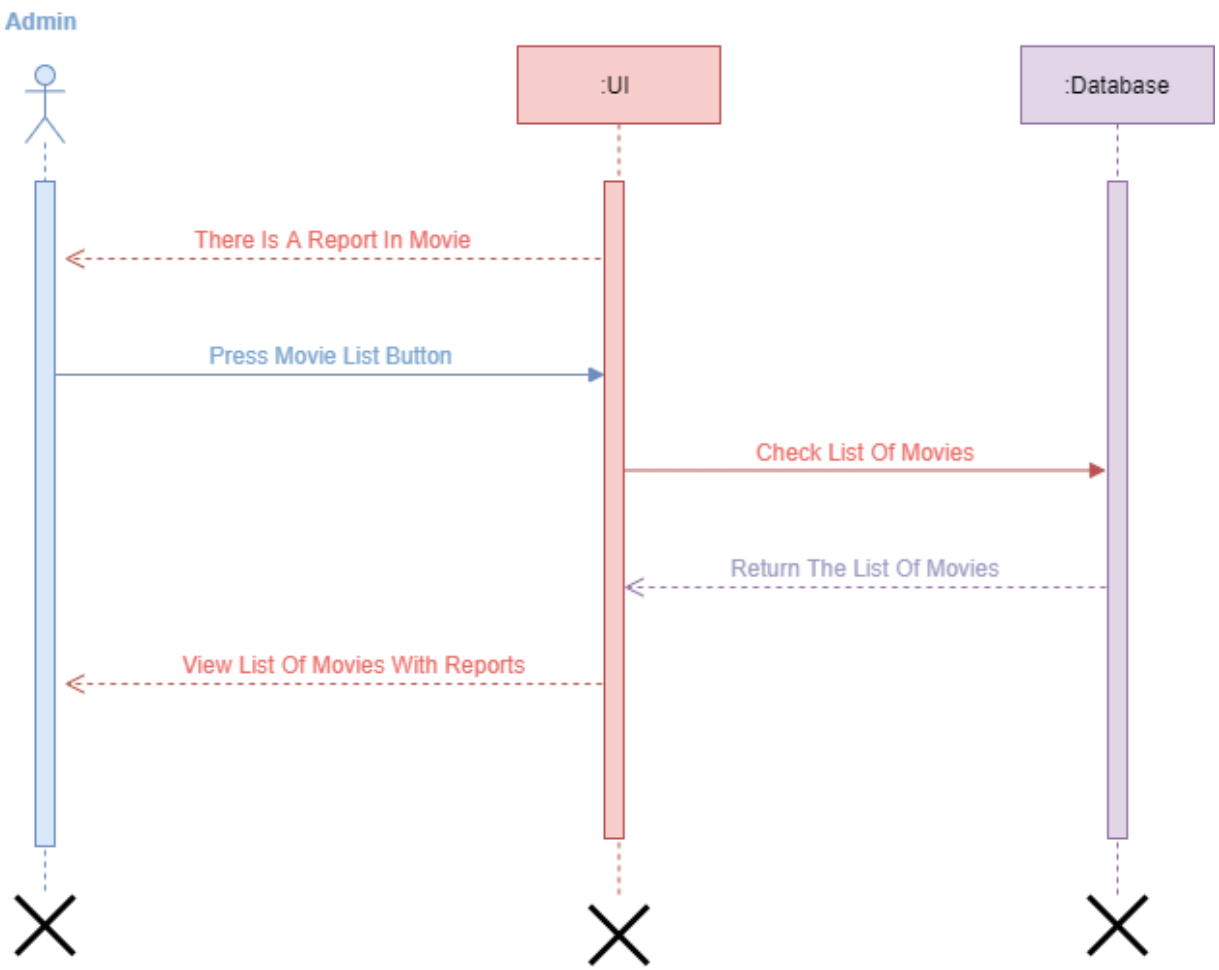
1- Use case diagram:



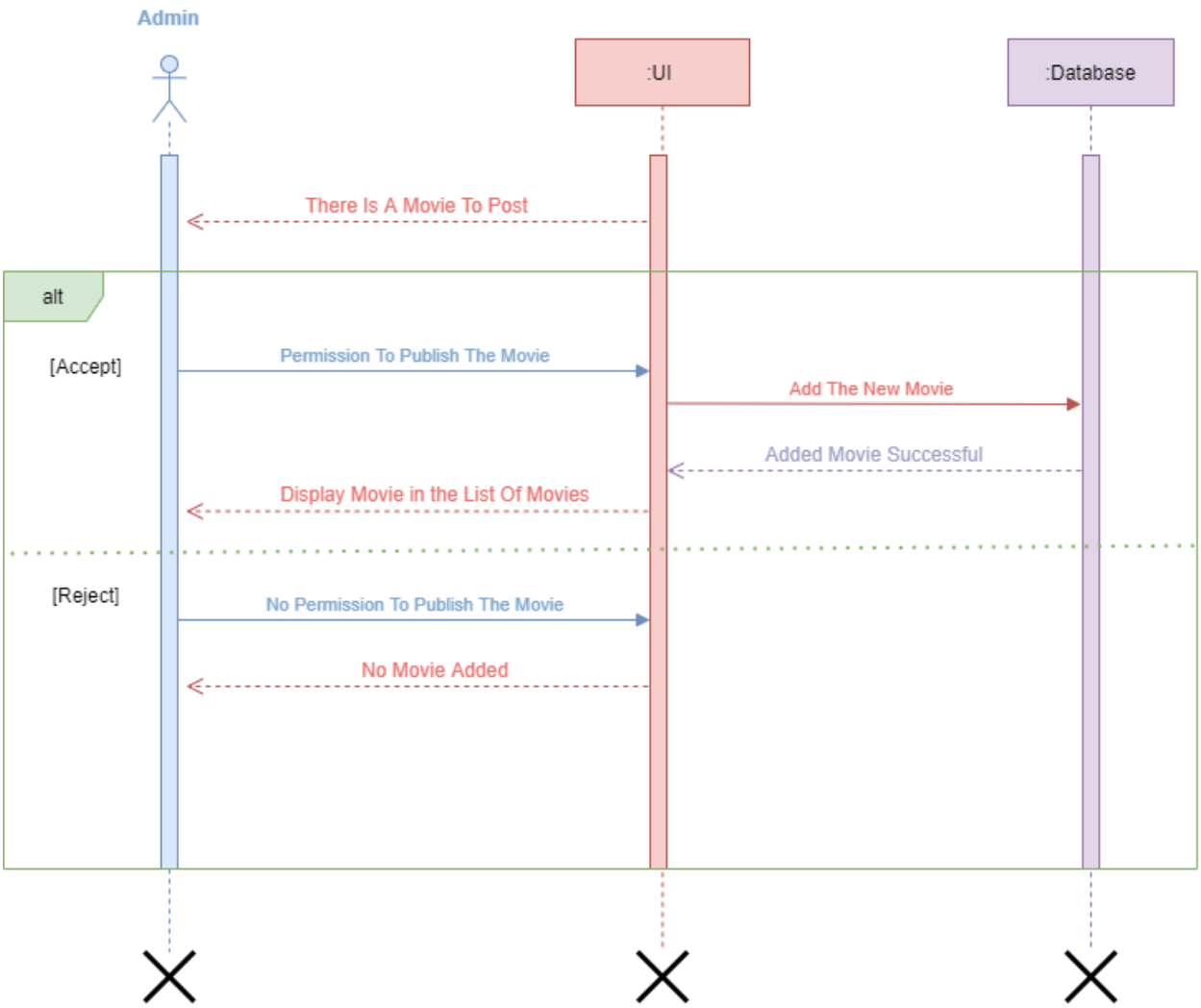
2- sequence diagram:

ADMIN

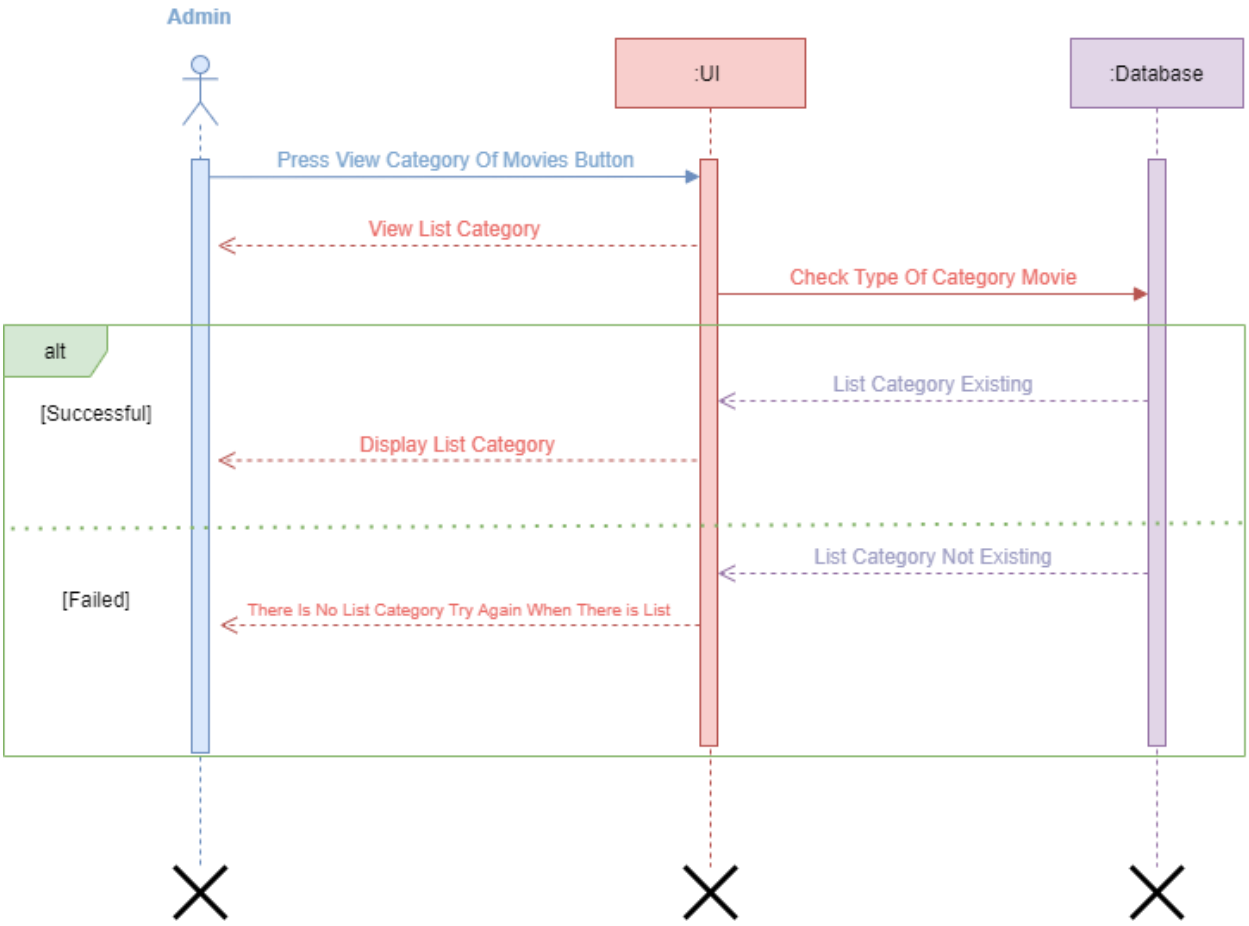
- Report:



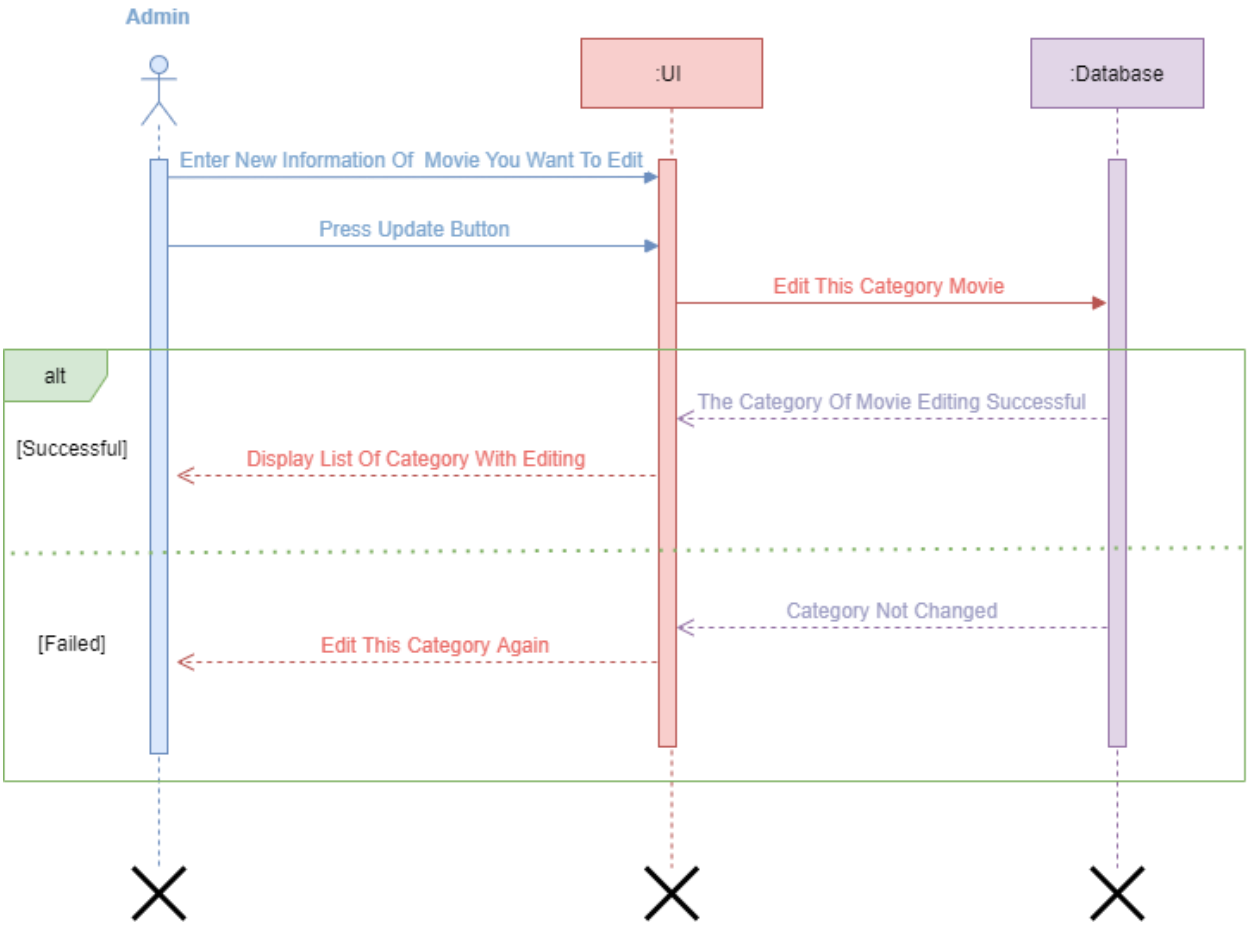
- Is Approved:


Admin
:UI
:Database
There Is A Movie To Post
alt
[Accept]
Permission To Publish The Movie
Add The New Movie
Added Movie Successful
Display Movie in the List Of Movies
[Reject]
No Permission To Publish The Movie
No Movie Added

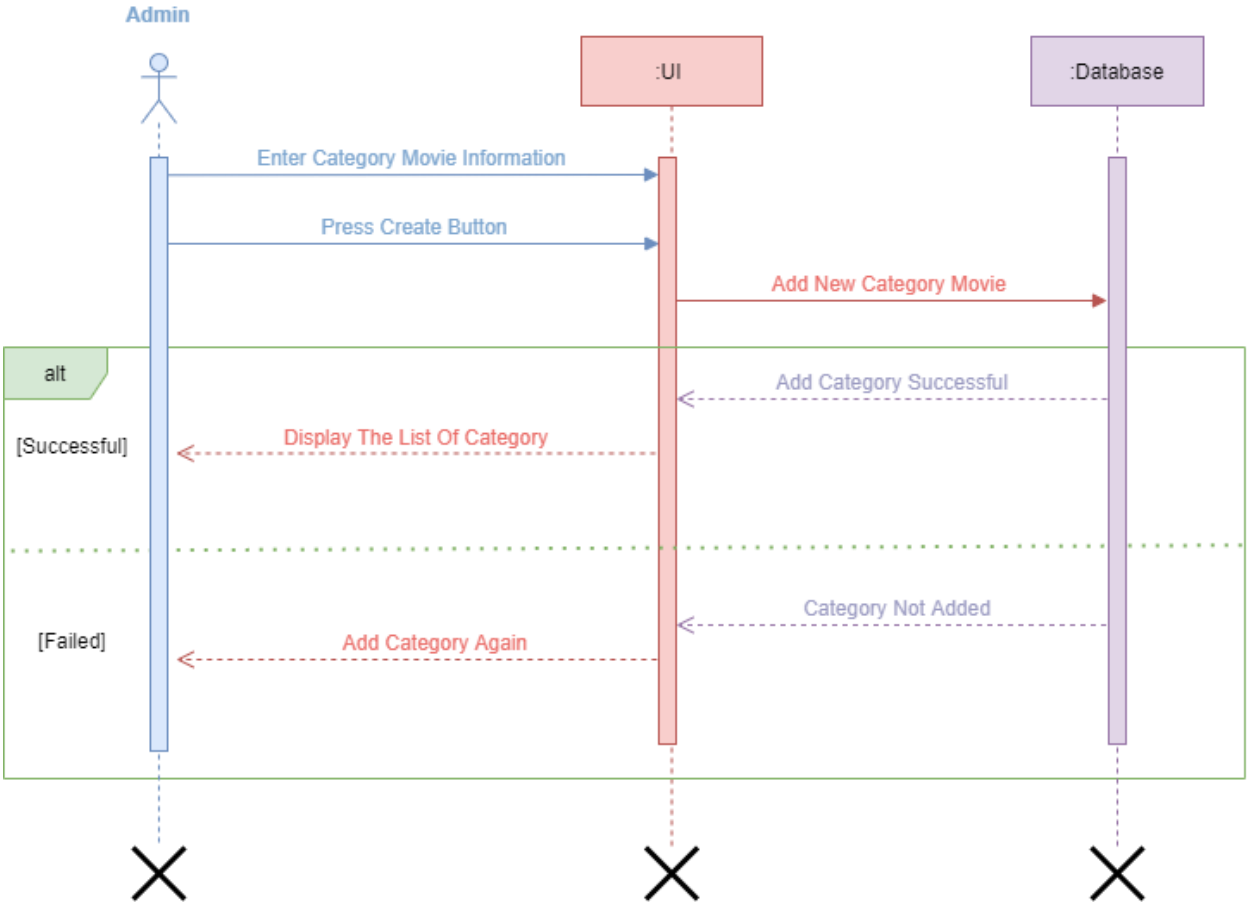
- List Category:



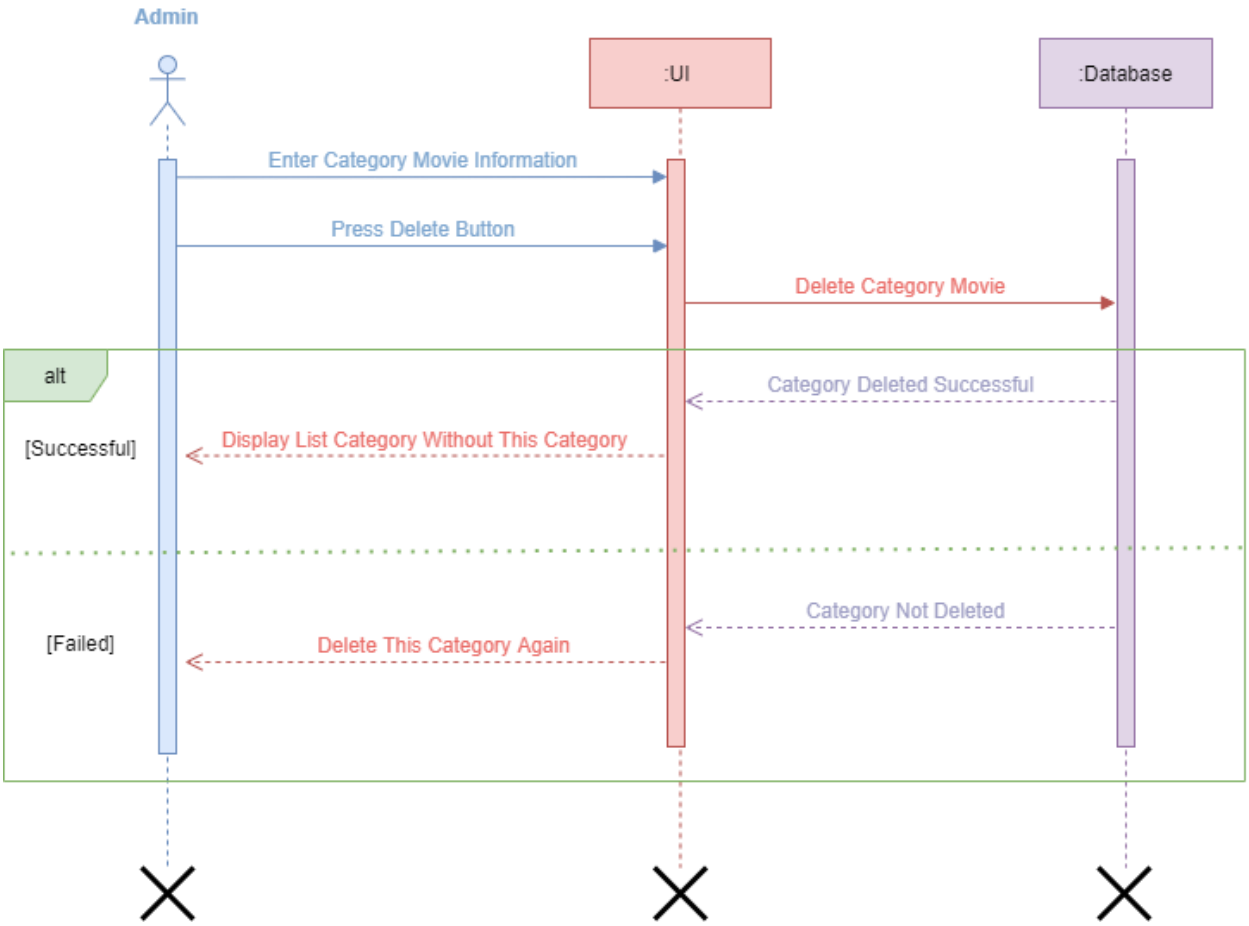
- Update Category:



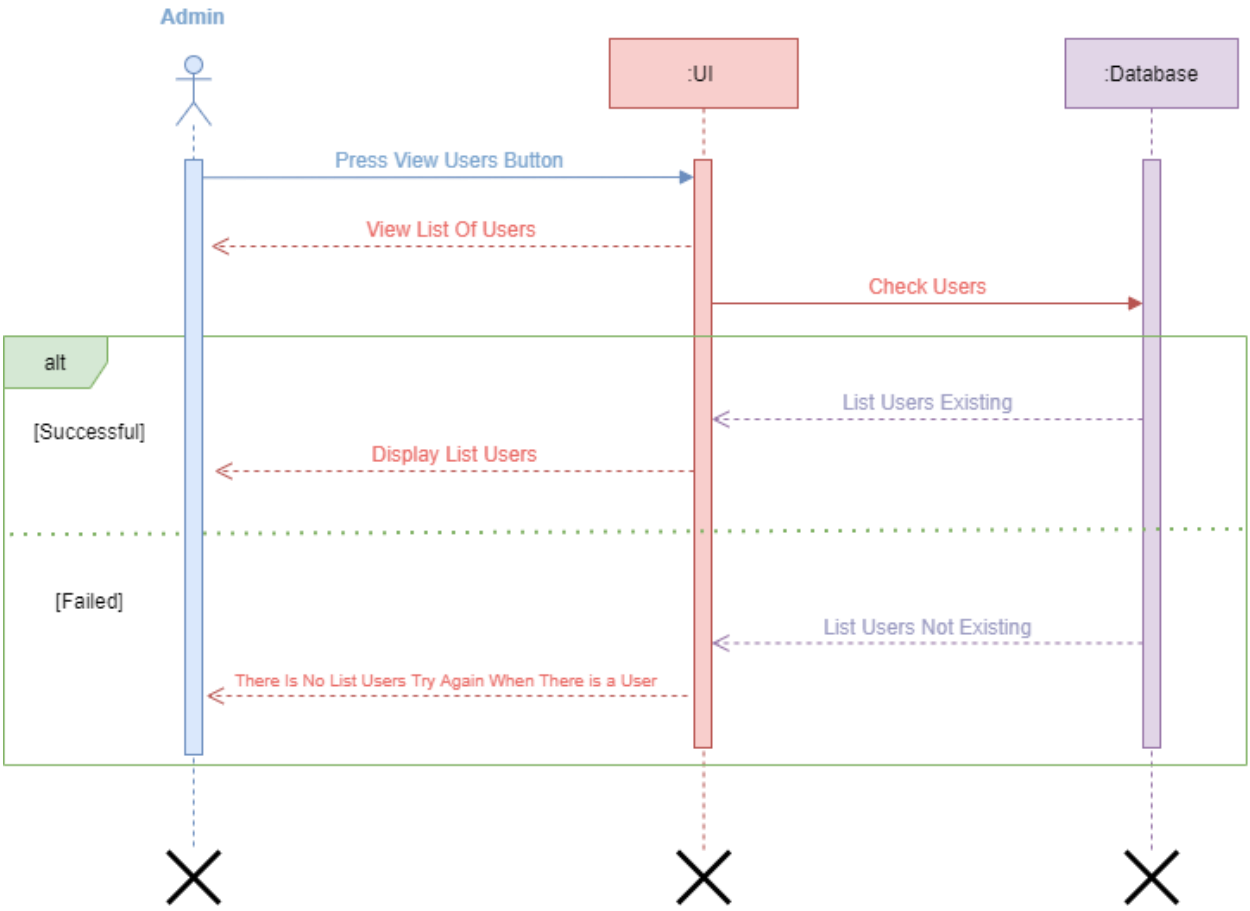
- Create Category:



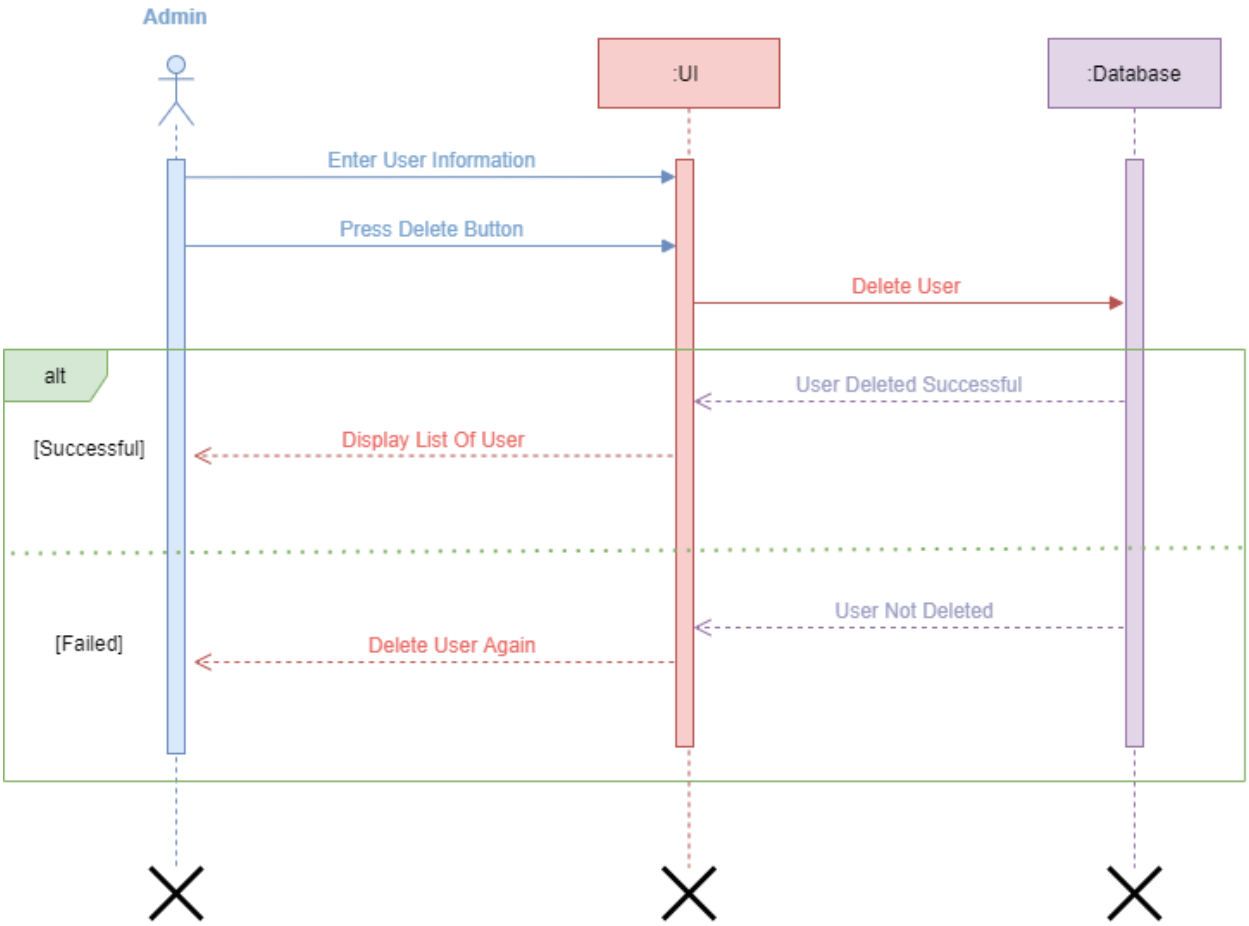
- Delete Category:


Admin
:UI
:Database
Enter Category Movie Information
Press Delete Button
Delete Category Movie
alt
Category Deleted Successful
[Successful]
Display List Category Without This Category
Category Not Deleted
[Failed]
Delete This Category Again

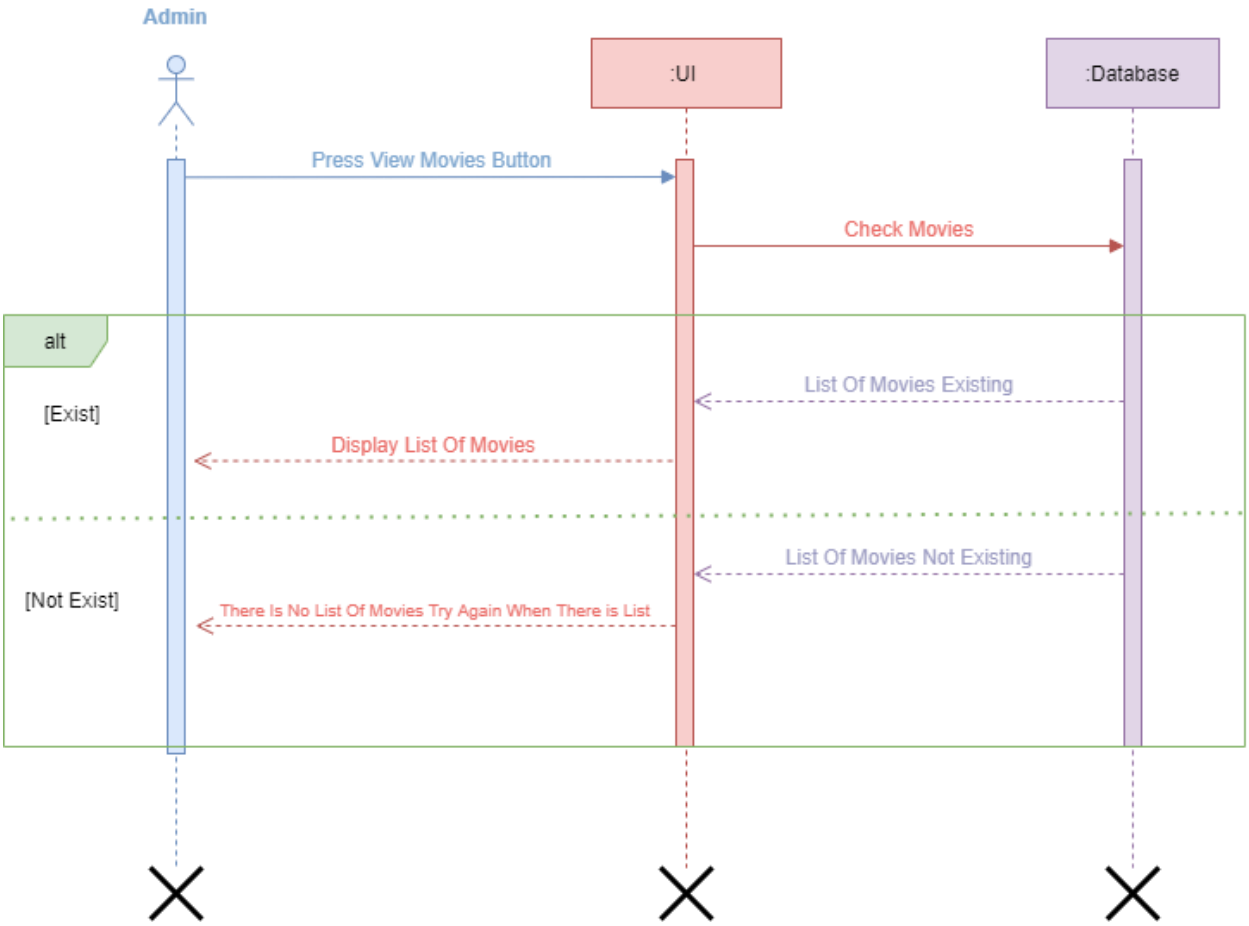
- List User:


Admin
:UI
:Database
Press View Users Button
View List Of Users
Check Users
alt
[Successful]
List Users Existing
Display List Users
[Failed]
List Users Not Existing
There Is No List Users Try Again When There is a User

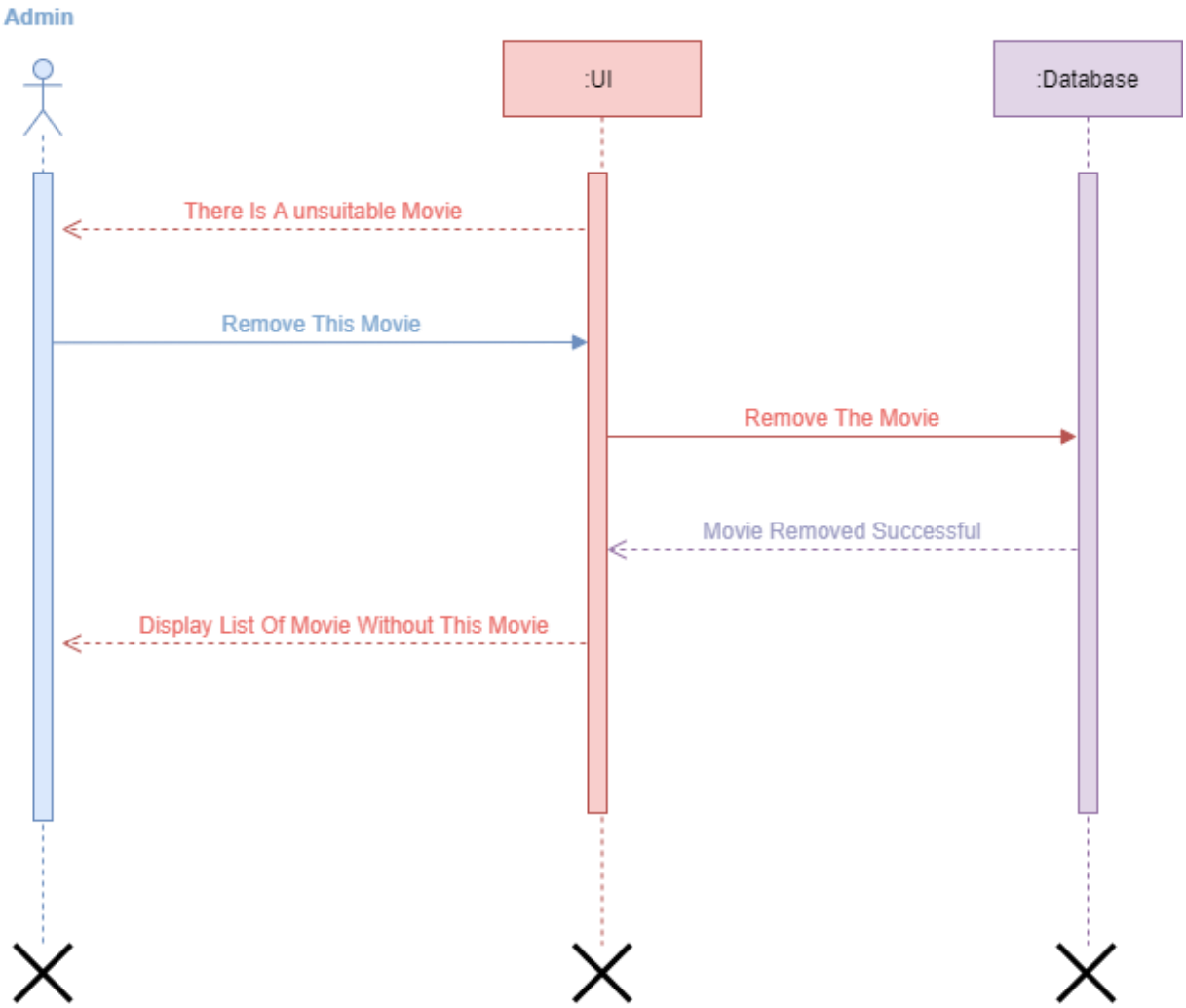
- Delete User:


Admin
:UI
:Database
Enter User Information
Press Delete Button
Delete User
alt
User Deleted Successful
[Successful]
Display List Of User
User Not Deleted
[Failed]
Delete User Again

- List Movies:

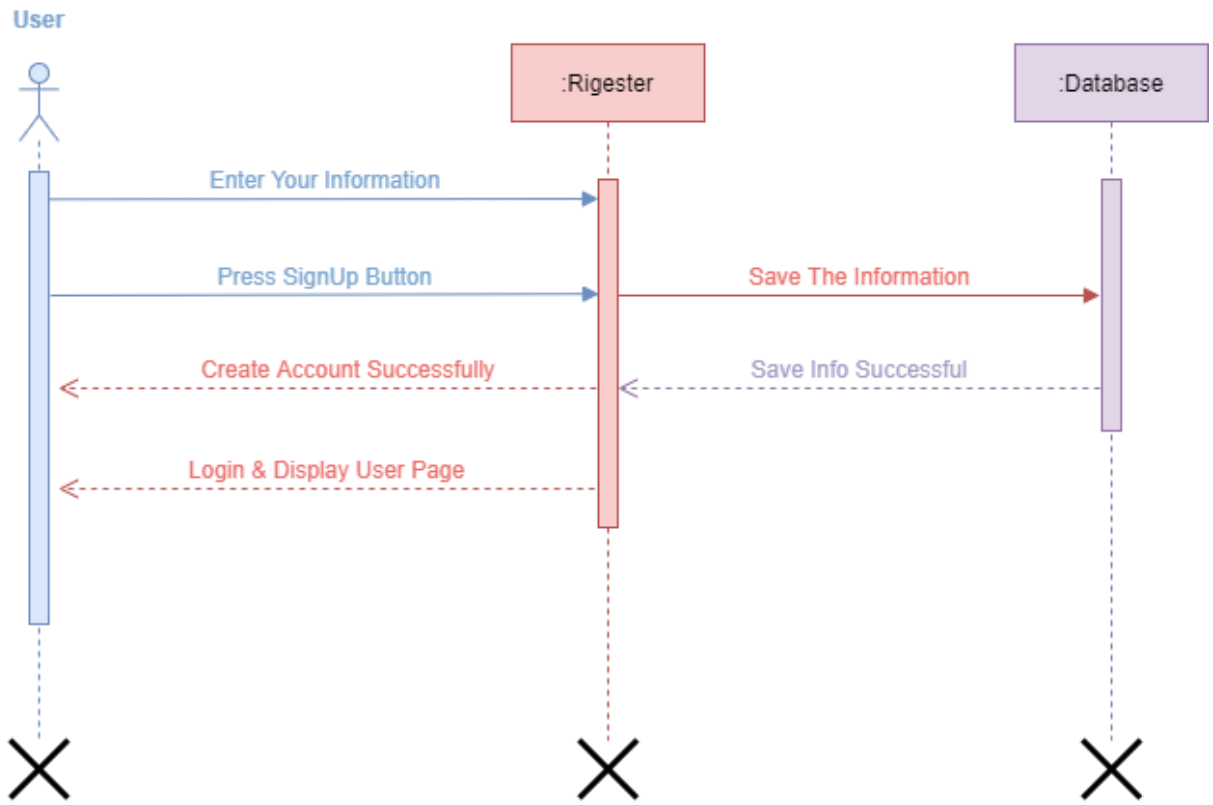


- Remove Movie:



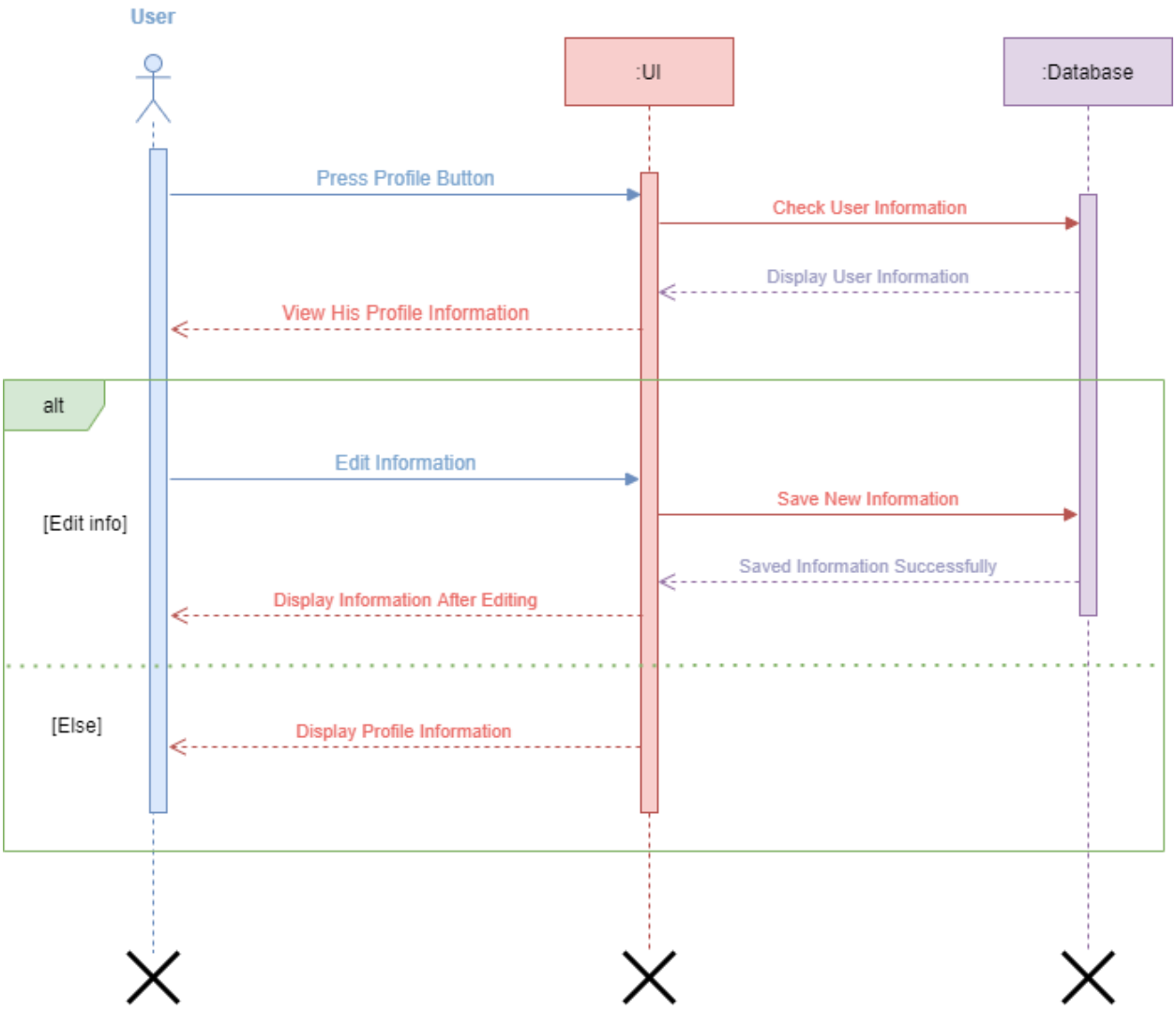
VIEWER

- Register:

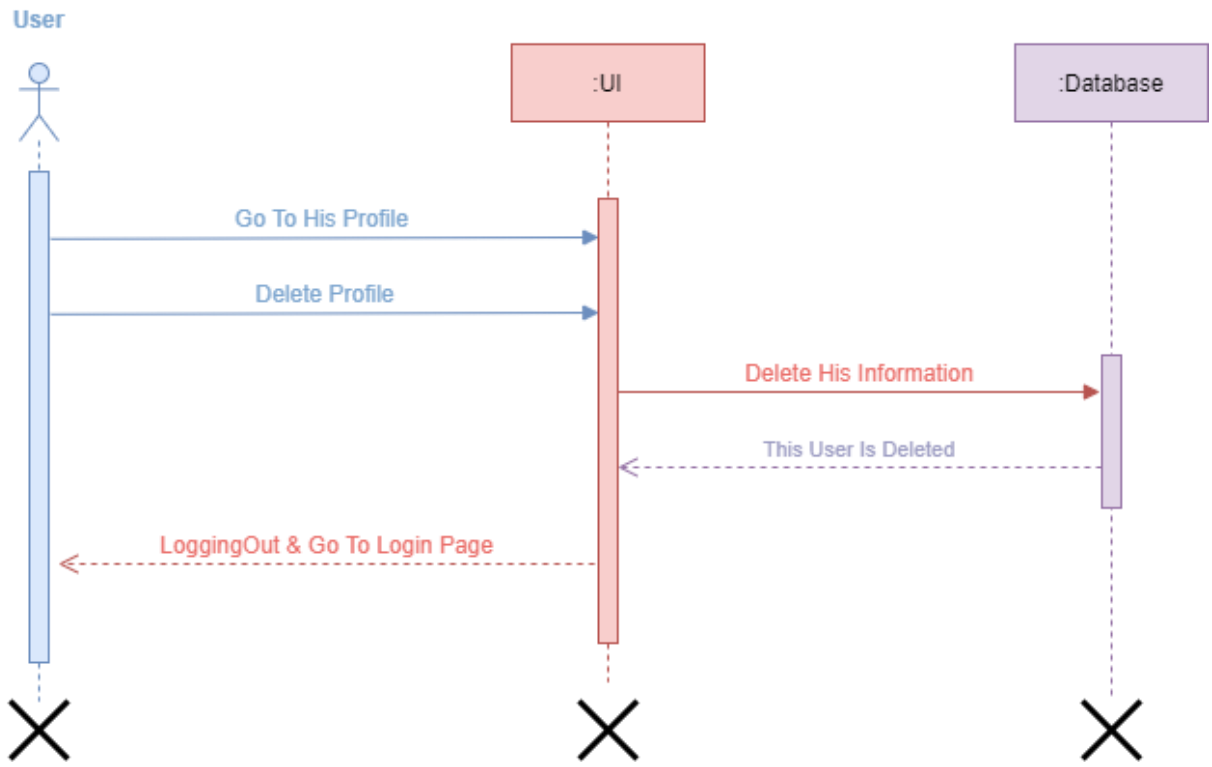


USER

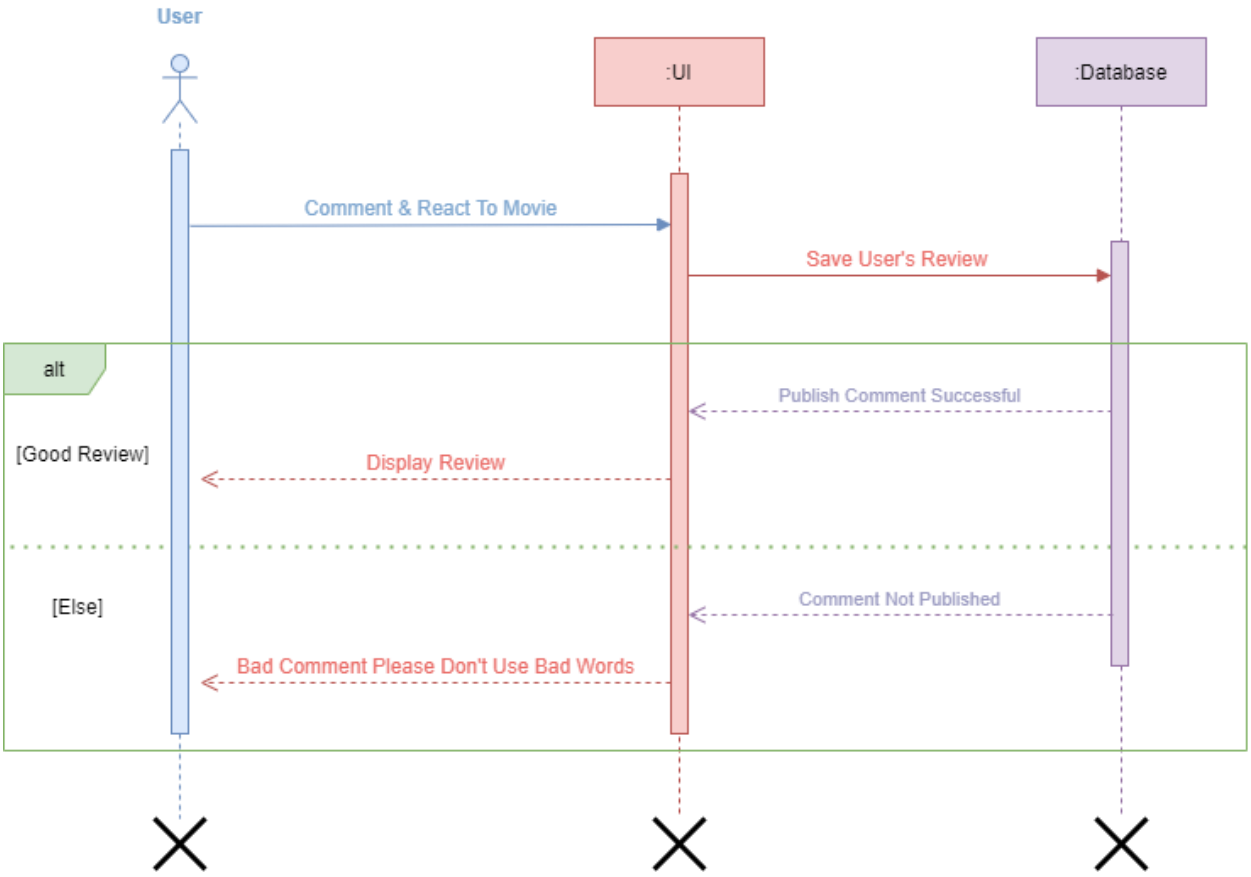
- Profile:



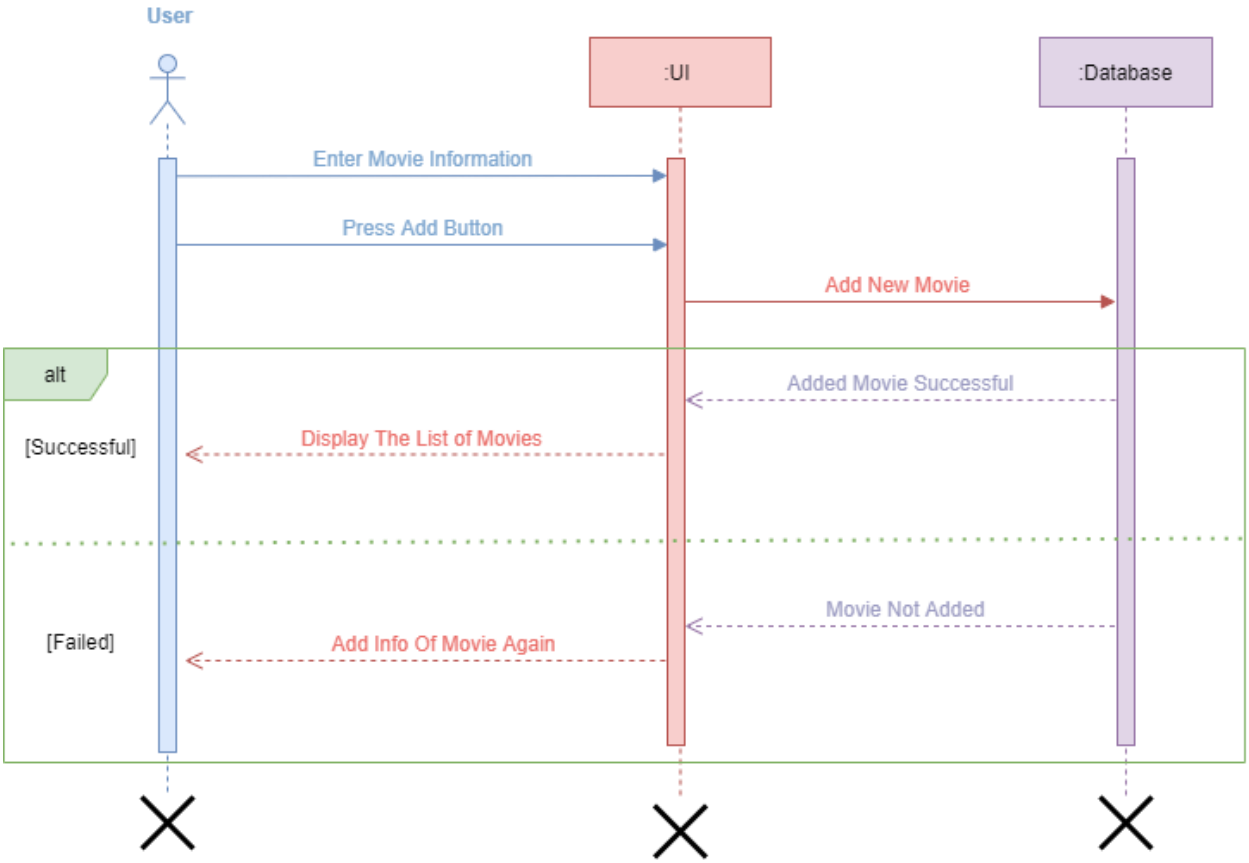
- Delete Profile:


User
:UI
:Database
Go To His Profile
Delete Profile
Delete His Information
This User Is Deleted
LoggingOut & Go To Login Page

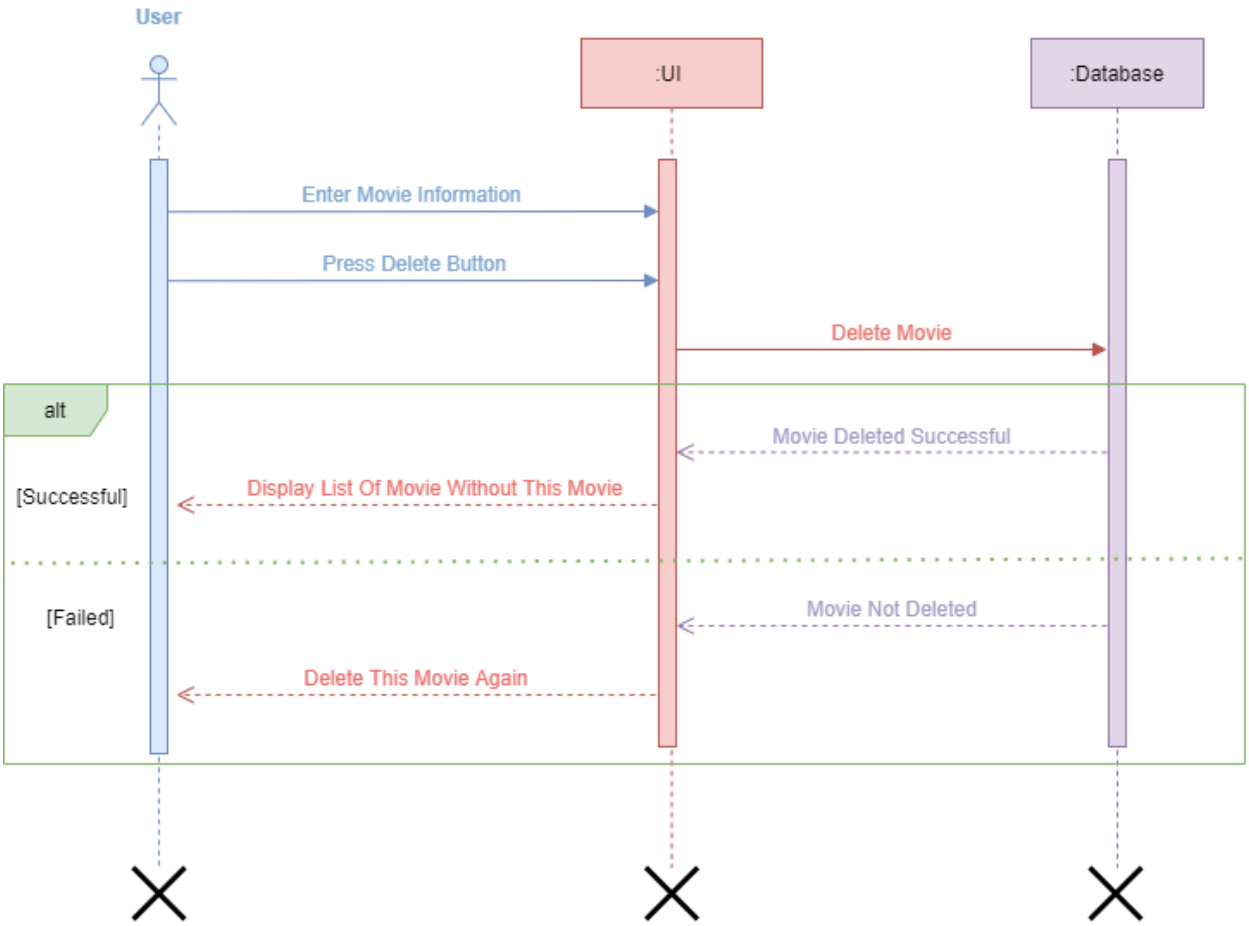
- Review:


User
:UI
:Database
Comment & React To Movie
Save User's Review
alt
Publish Comment Successful
[Good Review]
Display Review
[Else]
Comment Not Published
Bad Comment Please Don't Use Bad Words

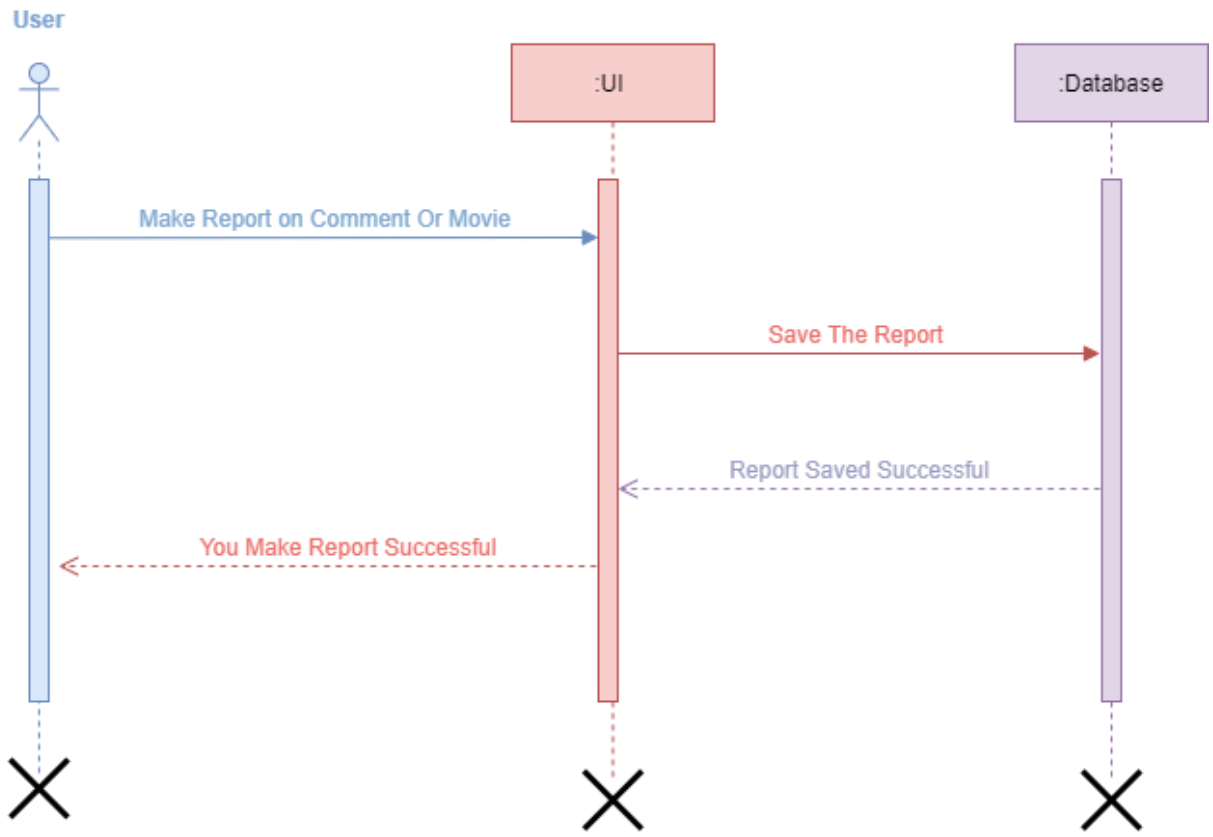
- Add Movie:



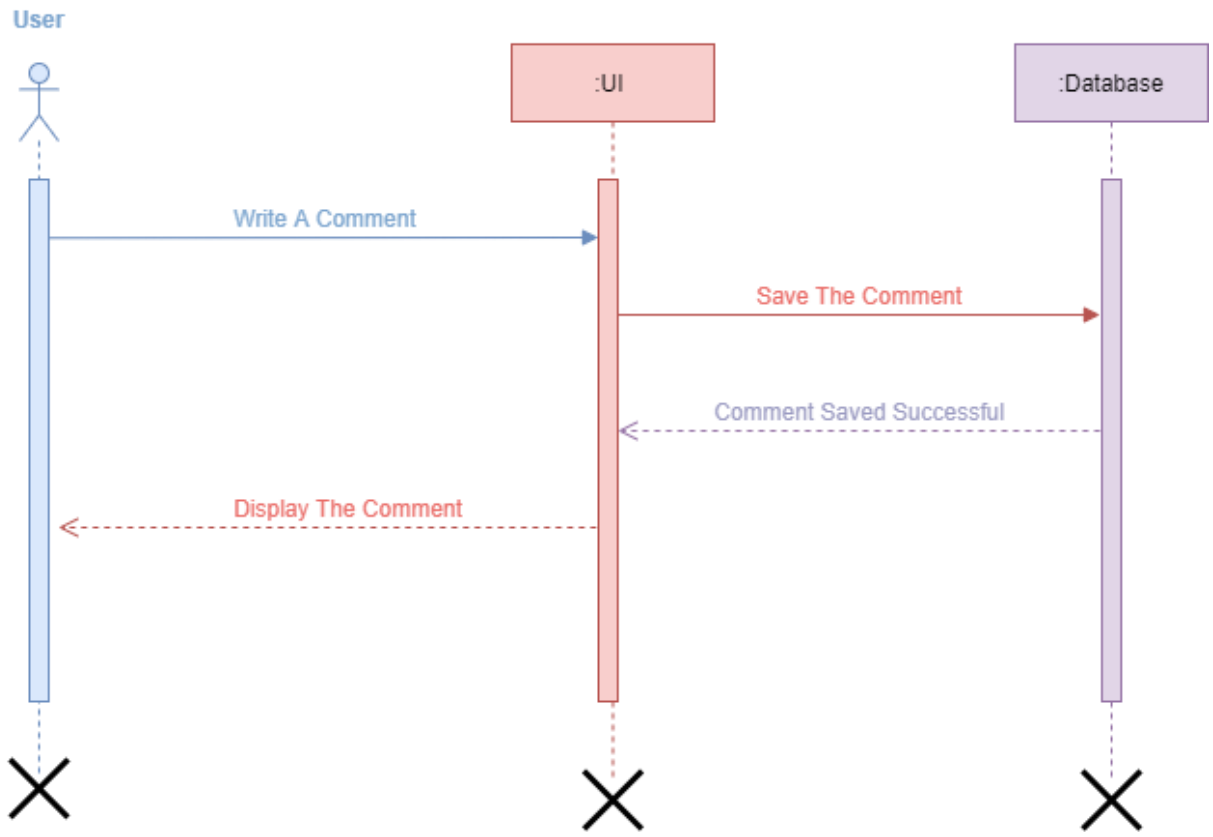
- Delete Movie:


User
:UI
:Database
Enter Movie Information
Press Delete Button
Delete Movie
alt
Movie Deleted Successful
[Successful]
Display List Of Movie Without This Movie
[Failed]
Movie Not Deleted
Delete This Movie Again

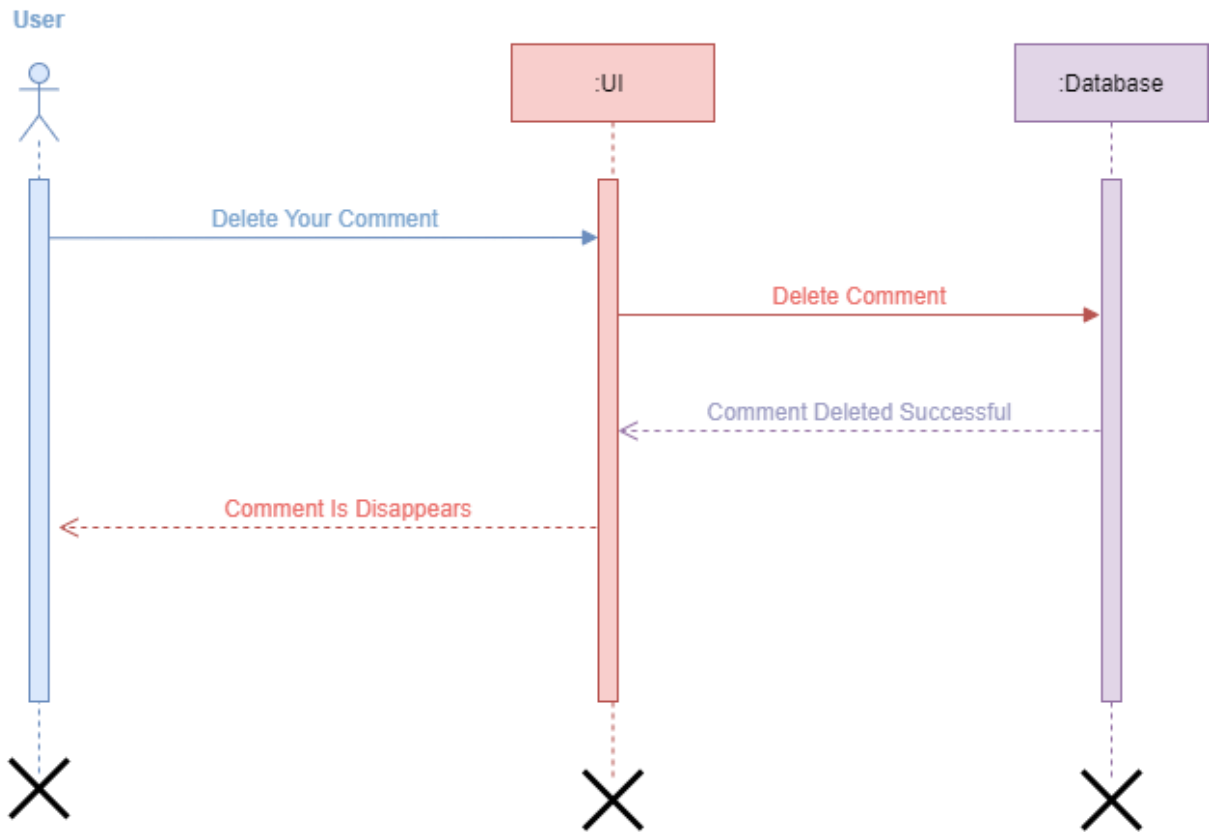
- Add Report:


User
:UI
:Database
Make Report on Comment Or Movie
Save The Report
Report Saved Successful
You Make Report Successful

- Add Comment:

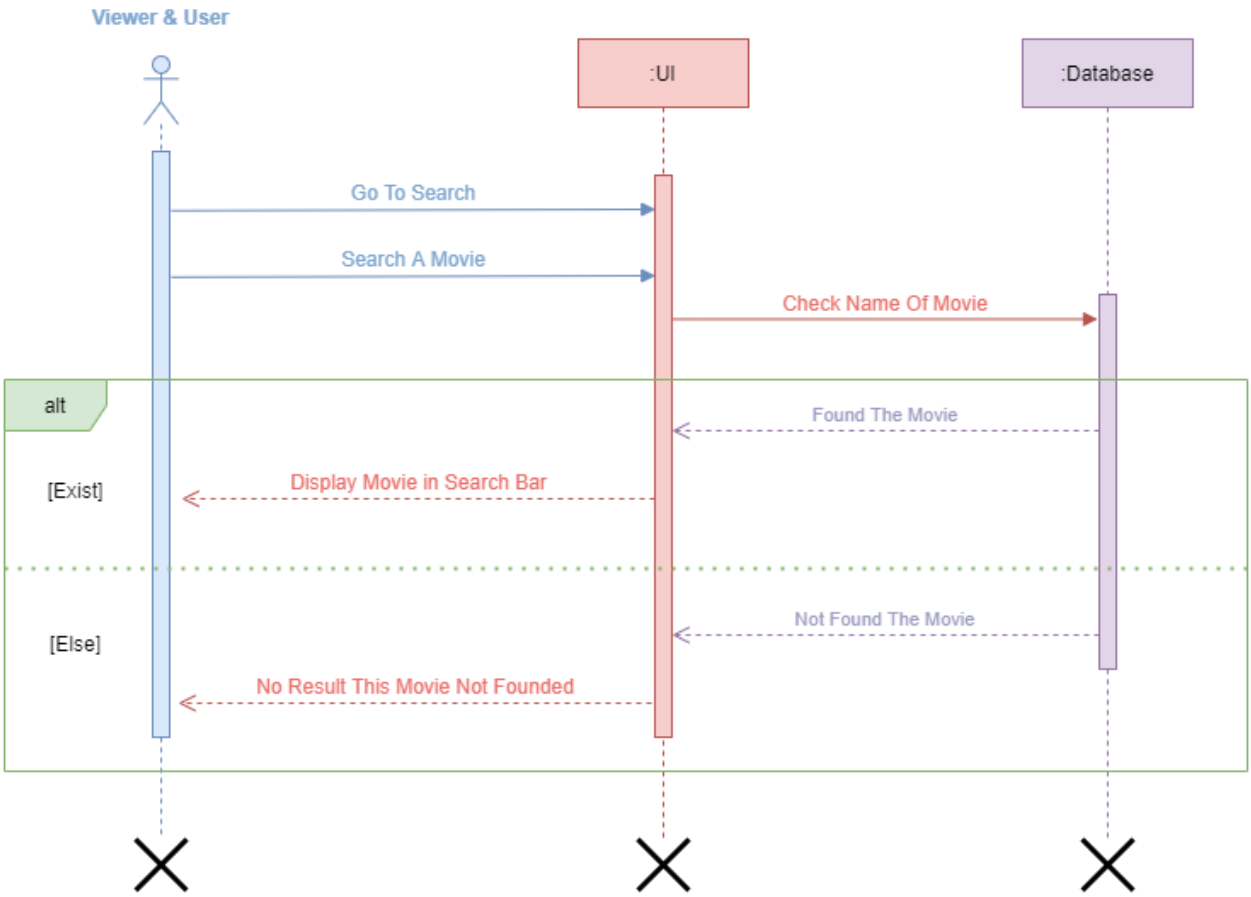


- Delete Comment:



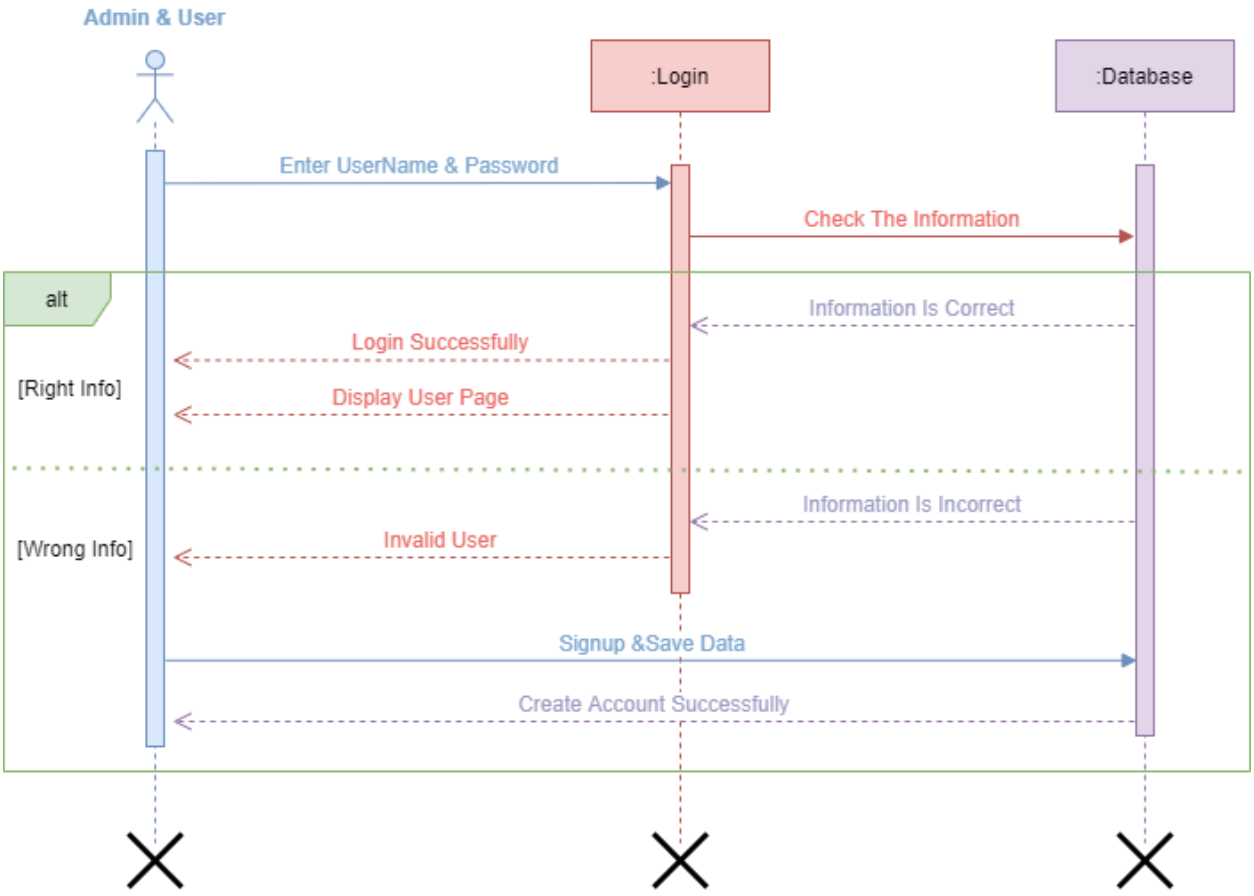
VIEWER & USER

- Search:

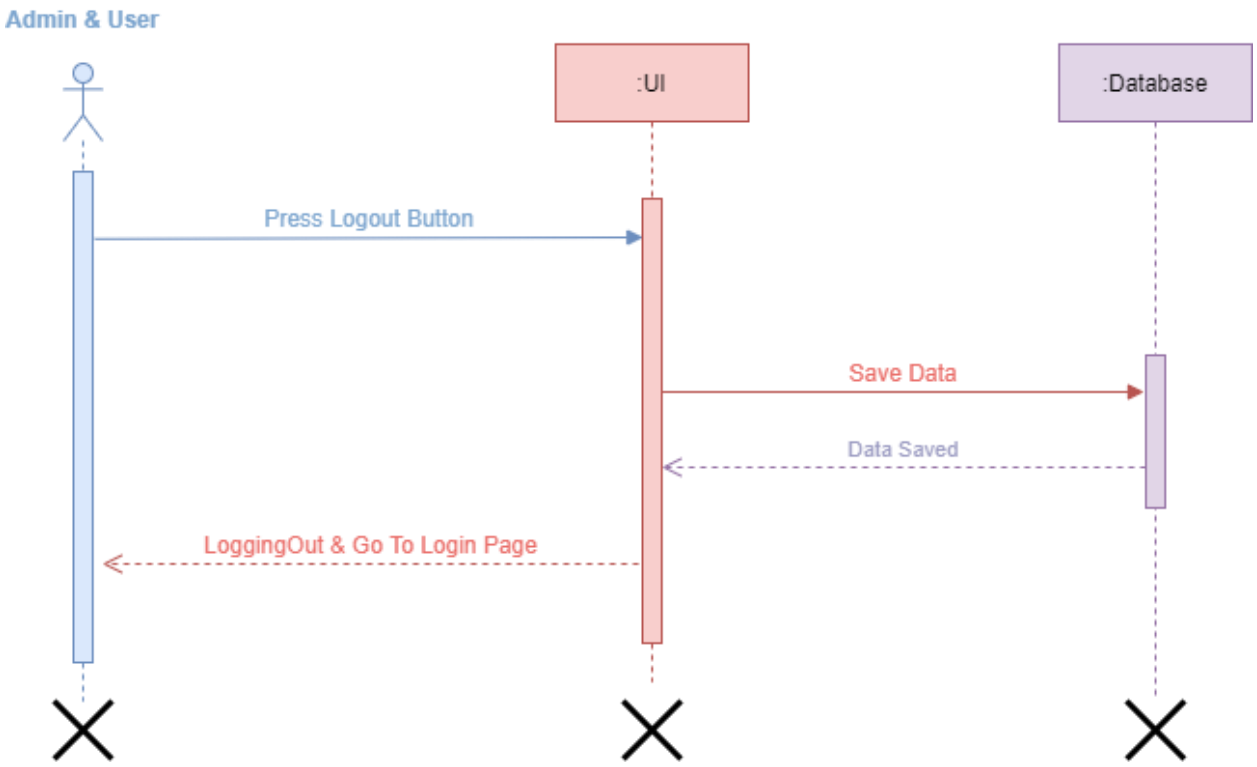


ADMIN & USER

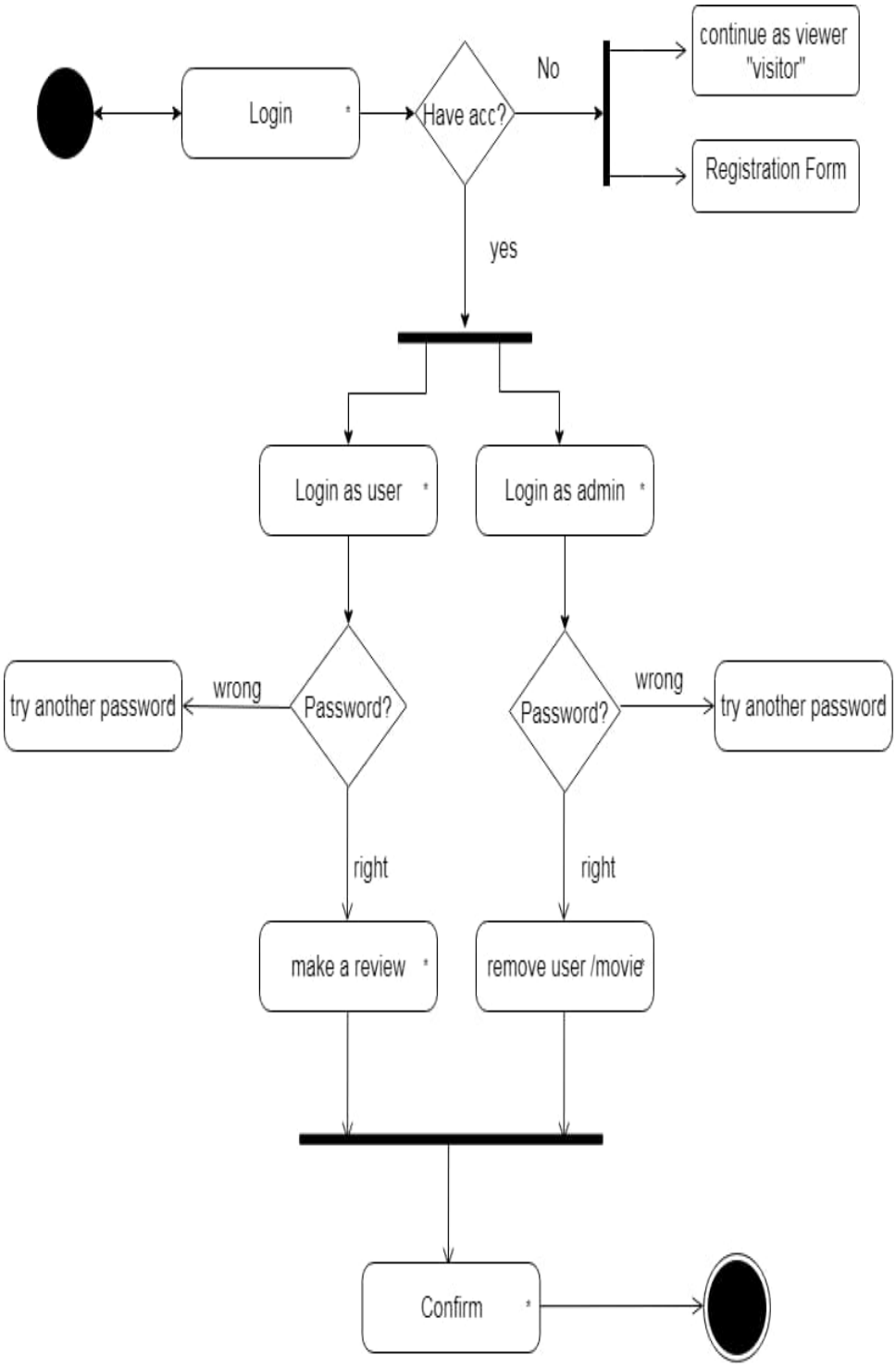
- Login:


Admin & User
:Login
:Database
Enter UserName & Password
Check The Information
alt
Information Is Correct
Login Successfully
[Right Info]
Display User Page
Information Is Incorrect
[Wrong Info]
Invalid User
Signup &Save Data
Create Account Successfully

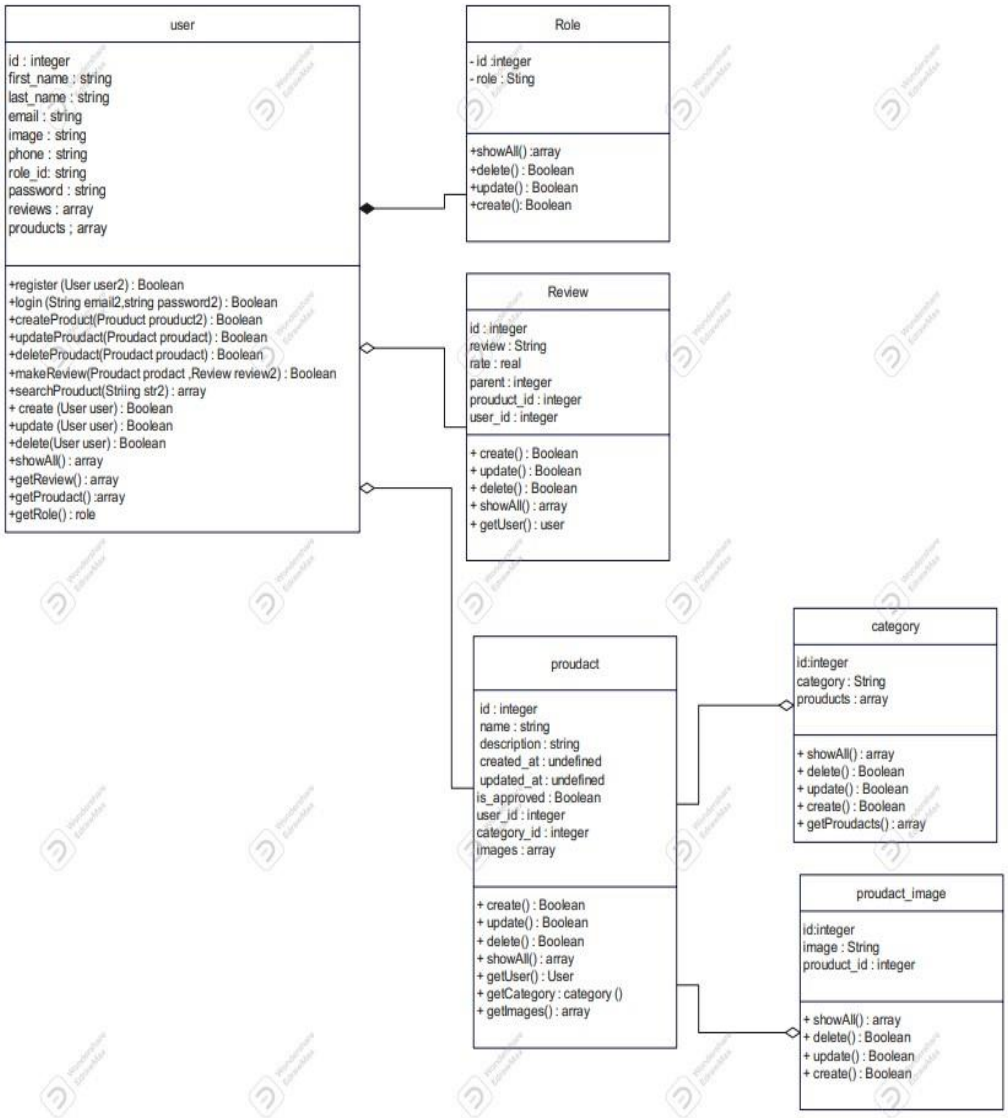
- Logout:


Admin & User
:UI
:Database
Press Logout Button
Save Data
Data Saved
LoggingOut & Go To Login Page

3-activity diagram:

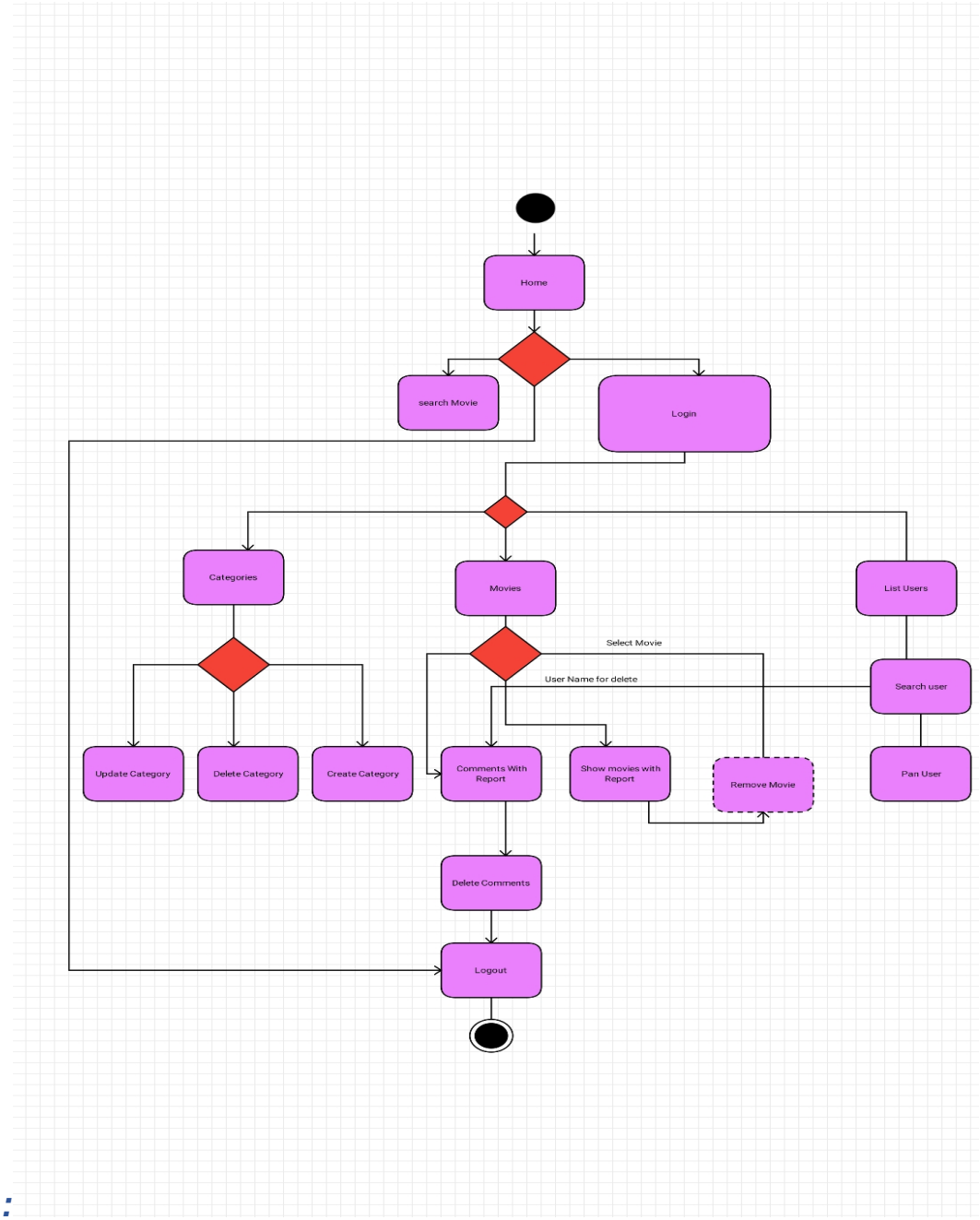


4-class diagram:



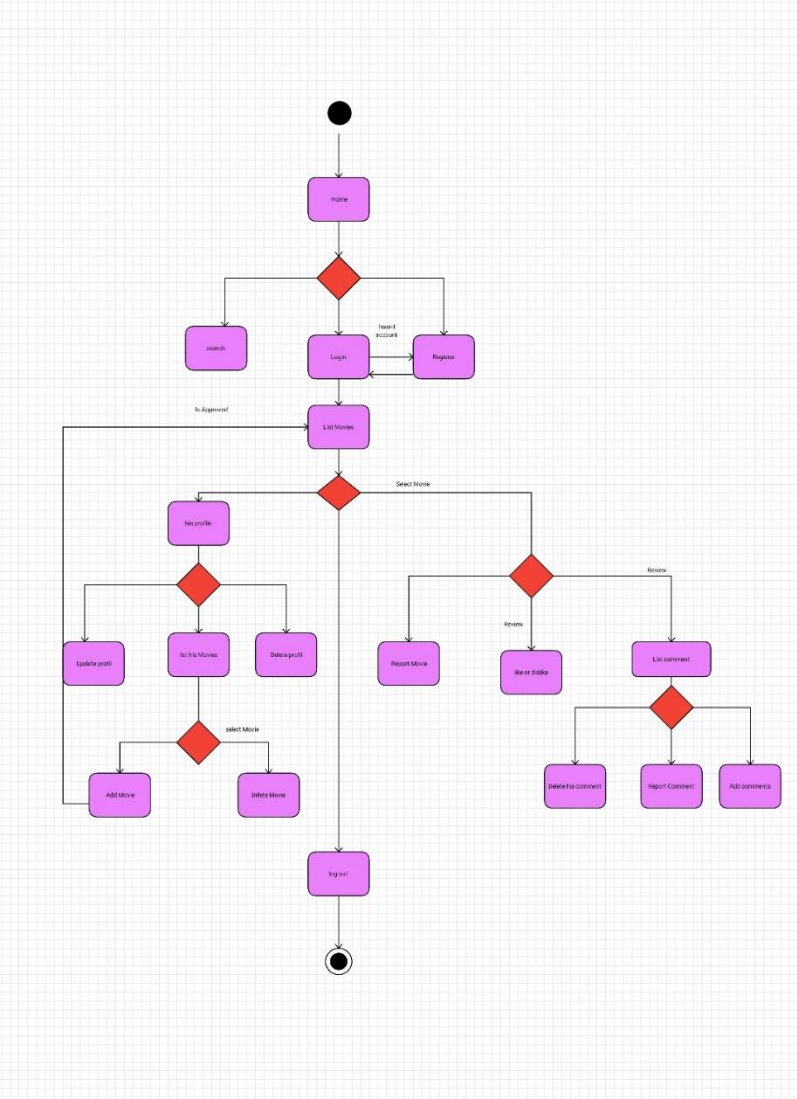
5- state machine :

Admin :



:

User:


Home
search
Login
Register
List Movies
Select Movie
his profile
Update profil
Delete profil
Report Movie
like or dislike
List comment
Add Movie
Delete Movie
Delete his comment
Report Comment
Add comments
log out

ERD:

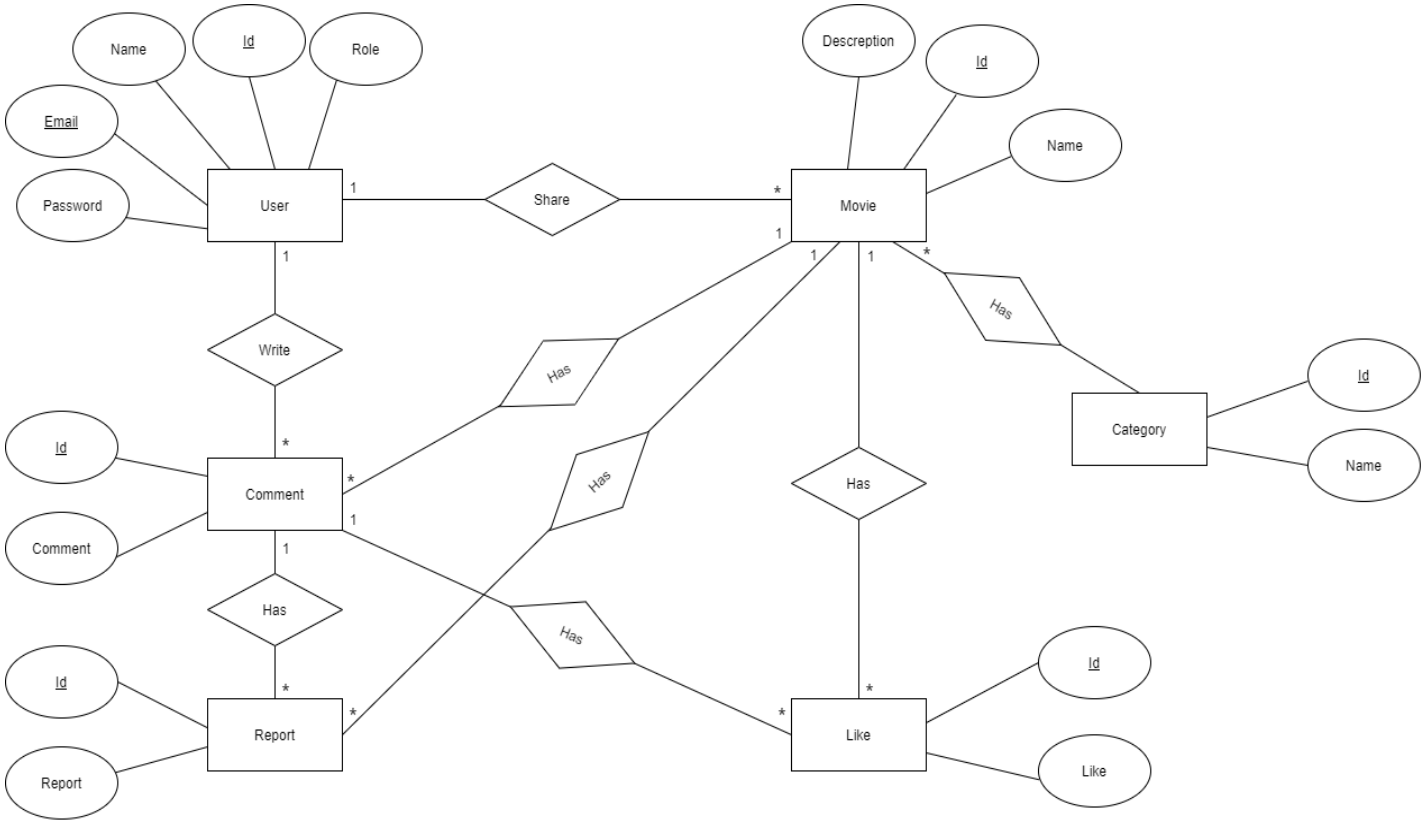
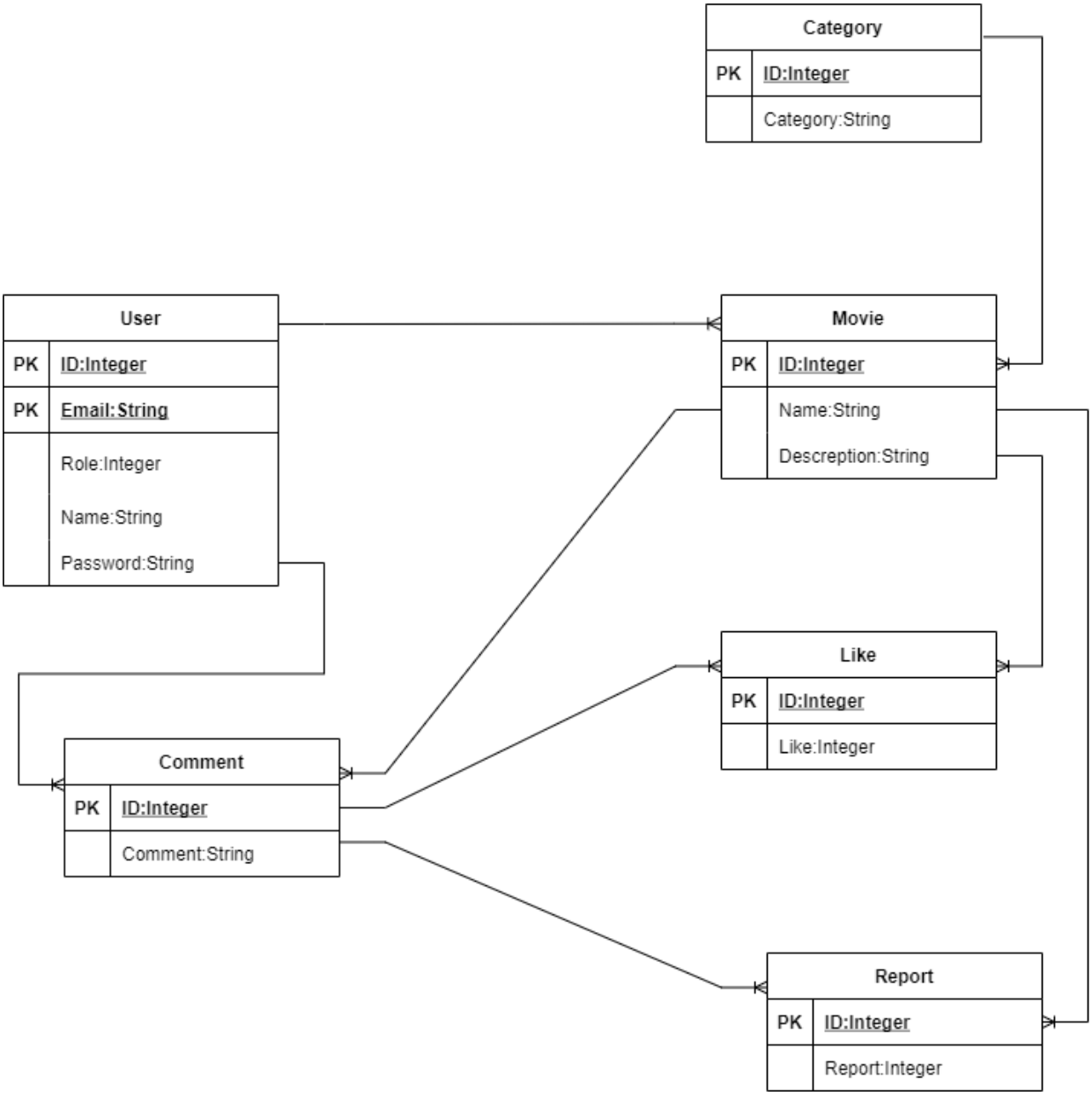

Name
Id
Role
Email
Password
User
1
Share
*
Descreption
Id
Name
Movie
1
1
1
*
Has
Id
Name
Category
Write
1
*
Has
Id
Comment
*
1
Comment
1
Has
Has
Has
Has
Id
Report
*
*
Report
*
*
Like
Id
Like

Table:


Category
PK
ID:Integer
Category:String
User
PK
ID:Integer
PK
Email:String
Role:Integer
Name:String
Password:String
Movie
PK
ID:Integer
Name:String
Descreption:String
Like
PK
ID:Integer
Like:Integer
Comment
PK
ID:Integer
Comment:String
Report
PK
ID:Integer
Report:Integer