Chapter 3 - Methodology

3.1 Introduction

Methodology is the section where the set of procedures are described in detail, providing the information regarding different rules and techniques that are going to be used to solve the research problem and achieve some specific result. It is a systematic process where collection of data and analysis of data is clearly stated with the well working plan by using different algorithms and machine learning techniques. Overall, it is an important step, where the research is conducted in order to increase the validity, usability and reliability of the obtained results.

The purpose of the recommender system is to predict and suggest the movies to the viewers. But for that all to have happen the system needs to be created. And in this system, we are going to use two different kinds of approaches to extract the relative features and provide the user with the most accurate output. The two approaches are going to be Bag of words model, Jaccard similarity and implementation of KNN respectively. The flowchart describing the overall process of involved in recommendation system is represented in Figure 3.1.

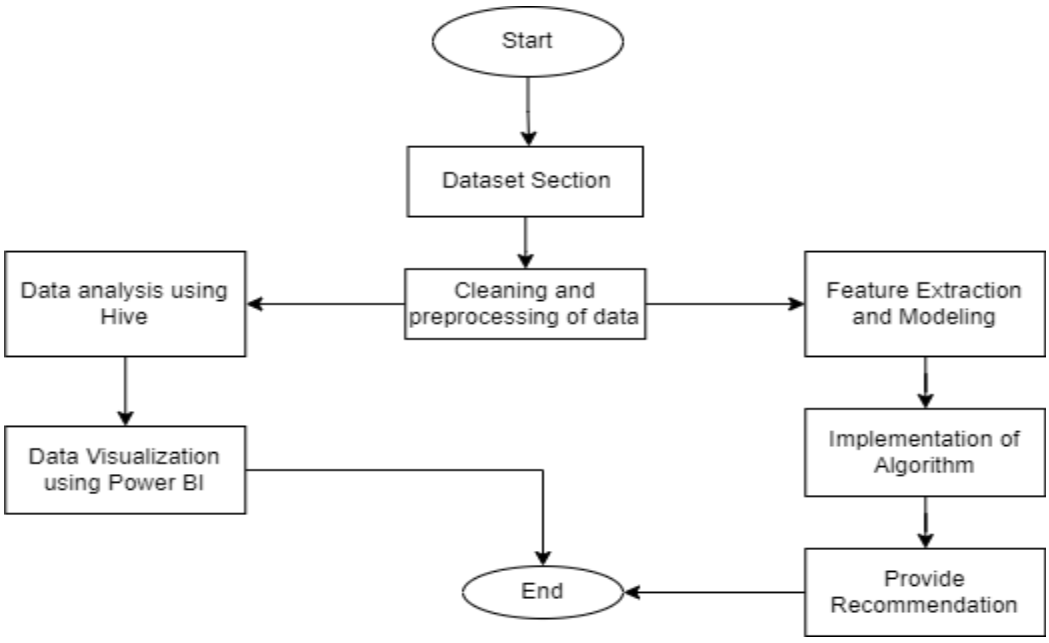

Figure 3.1: Flowchart describing the process of Movie recommendation system

For reminder, the aim of the proposed system is to provide the user with good personalized movie suggestion.

And the steps to achieve the goal, the objectives taken are

- To help discover new features in movies and better understanding of data.
- To handle large amount of data related to movies and extract meaningful information.
- To overcome the problem of sparsity, which is common in collaborative filtering system.
- To suggest the movies to an individual of their interest using content-based algorithms.

- To make the successful recommendations and analysis by using big data technologies.

As it works with words and meanings and the implementation will be experimental so, it is qualitative approach. The data will be analysed using different techniques of coding and thematic analysis, where the relationship among the data will be discussed by identifying the pattern and themes between the data. The strength of qualitative approach is such that it can work with detailed data providing rich information and is able to provide the result which cannot be obtained through quantitative approach. But it sure does have some limitations such as it is hard to generalize large amount of data, the process can be time consuming and intensive.

3.2 Tools Used

3.2.1 Excel

Microsoft Excel is the software program which allows users to create, analyse, manipulate data and performs calculations. It has many other users and different organization use it for different purpose. It is versatile, efficient and powerful tool allowing users to manage and analyse data efficiently. In this project Excel has been used for better understating of data, finding the irregularities among the data, cleaning the data and basic analysis of data. (Gillis, 2021)

3.2.2 Python (Google Colaboratory)

Python is one of the easiest and most powerful programming tools used in wide range of applications including machine learning, web development and data analysis. Google Colab on other hand is a cloud-based platform, which creates the environment for users to run python code on it. In this project python programming language is used to pre-process the data, feature extraction and implementation of machine learning to perform the analysis of data and to generate desired output. (Alba, 2022)

3.2.3 Hive and Hive Query Language

To run Hive and Query language the environment needs to be setup using virtual machine (VirtualBox), it allows multiple system to run on it, it creates virtual machine to run Hadoop and other big data tools, it has two main components Hadoop file system and MapReduce which allows processing of large datasets, data storage and analysis of data. In Hadoop there is pre-configured virtual environment that has certain tools and application known as sandbox. It includes Hive, Pig, Spark and so on. In this project Hive is used to import and store data, and with the help of Hive Query language the analysis of data is carried out to give insight of data. (Harshitn, 2022)

3.2.4 Power BI Desktop

Power BI is an analytical tool used in development of business. It is used for creating report, visualizing data, analysis of data and to give the life to data. It is useful in providing real-time analytics and share information through interactive dashboard. In this project Power BI is used for visualization of data, providing the clean and pre-processed data, it is able to provide the result in different forms of charts and graphs. (Scardina & Horwitz, 2022)

Machine learning

In relevant terms learning from the data and improving the performance of system over the period of time with better dataset, which involves training the data to the machine, feeding the features into the machine and fitting it in the model, generating different pattern and

relationships between the data to make decision and prediction for the future is known as machine learning. There are different kinds of machine learning algorithms some of them are supervised learnings, Unsupervised leaning, Semi-supervised leaning and reinforced leaning. Every machine learning algorithm has been categorised based on their working mechanism. It has huge potential and huge practical applications, such as image processing, natural language processing, recommendation system, speech recognition, fraud detection and so on. Not just in the filed of data science but it can also be used if different sectors including business, arts and so on.

Big Data Technology

Big data technology contains different tools and techniques, which are used to process, analyse, and store huge amount of data sets. Apache spark is one of the most famous and widely used big data technology. It is an open-source distributed computing system which is capable of handling large-scale data. Spark provides the environment and programming interface, on other hand pySpark acts as API for spark, which allows the developer involved in python to enhance their power to process and handle large amount of data. Spark works with pySpark enabling different pythonic libraries tools to be used like NumPy, Pandas, matplotlib and so on.

3.3 Execution of the project

3.3.1 Data Collection

In order to create a system or run a system we have to process the requires data. Data collection was quite an easy task for me as there are huge data for my projects present online. But according to my project requirement. I choose the dataset from Kaggle website I.e, (movie database) TMDB 5000 movie Dataset. Where dataset is quite simple and can be easily manipulated. It basically contains two datasets one regarding the information related to movies and other one being the dataset with credits of movies.

Movies dataset (budget, genres, id, keywords, language, title, overview, popularity, production company, revenue, runtime, release date, status, tagline, vote average, vote count)

Credits dataset (title, movie id, cast, crew)

The above dataset consists of both numerical and categorical data, but as the topic suggest the recommendation is basically about the contents in the movies, the recommendation is preformed based on different contents related to movie plots, structures, cast, crews and so on.

3.3.2 Data Cleaning and Pre-processing

This is the first step after collection of data where the data are processed and cleaned to get the required data set. Which can be used in actual system to train and test to get the accurate result. This involves various steps starting with importing the data set and second of all the duplicate and non-useable/relevant observations are made to remove the distraction and to maximize the output.

Next step would be to remove the errors from the data set which may have been their while importing the data. There may be outliers in the data which needs to be handled to increase the performance of system.

Likewise, the missing data needs to be handled, there are two ways of doing so the first option would be to drop the observation. Second would be to input the missing value without losing its integrity.

After all that, the data need to make sense. It needs to be in the specific format and should be useable in the system.

3.3.3 Modelling and Feature Extraction

Bag of words model is one of the best techniques for processing the text data and feature it as numerical vectors. Here the texts or sentences are represented as bag of words ignoring order or grammar but keeping the count of each word. To create the bag of words, several steps are followed starting with,

Tokenization: It is the process of splitting the sentences or texts into words or otherwise known as token.

Creation of Vocabulary and counting: The second process involves creating a vocabulary of unique words from the texts and counting the frequency of words

Vectorization: Finally, the text is represented as vector of word frequency.

In this system Bag of words model is used for feature extraction process from the text. In which the collected data is converted into list which is then converted into document vector removing all stop words, signs and punctuation. Then the words are scored and made ready for the further process. The vectors produced by the results are very high-dimensional. It's one of the most powerful but simple approach but it sure does have some limitations, it loses its context as a whole text or sentence unable to capture the essence and meaning of it.

Jaccard similarity is the part of machine learning technique, which is very useful in finding the common features in data and can also measure the distance between the datapoints, proving it to be one of the best algorithms for content- based data. In machine learning algorithms, it is important to measure the similarity or distance between the data points. Jaccard similarity can be defined as the part of intersection of two sets which is dived by the size of union of given two sets.

Jaccard similarity = (number of common features) / (total number of distinct features in both items)

For instance, Jaccard similarity= 3/6=0.5

That basically means if Jaccard similarity is 0 then there are no common features and if the value is 1 then they share the features in all their items.

In Jaccard similarity we measure the similarity between two sets of vectors. Let's consider there are two vectors doc1 and doc2 which is represented by the formula 1.

$$Jaccard\ Similarity(doc1, doc2) = \frac{doc1 \cap \mathrm{doc2}}{doc1 \cup \mathrm{doc2}} \quad \text{----------1}$$

After implementing the described process, we can now move to the prediction process in which the Viewer will search for the specific movie and based on that movie we will be able to recommend the nearest and closet movies based on his/her search. The movies recommended will be quite similar to the user preferences.

Content-based filtering is a recommender system based on the similarity between contents and items. It is one of the most used techniques to find the relationship between the attributes. In order to find similar item which matches users' interest Jaccard similarity is used which allows finding similarities among two sets. The different steps involved in process design of the movie recommendation system are listed below:

Output: Recommended movies

Step 1: Users inputs query (movie name) in the search section of the system.

Step 2: User query and dataset are pre-processed for extracting their feature.

Step 3: Pre-processed data are passed to feature extraction process using Bag of word model

Step 4: Extracted feature are passed through Jaccard similarity algorithm.

Step 5: Ten nearest movie with user query are compared.

Step 6: Ten common featured movies to user query are predicted and displayed as output in user interface.

3.4 Analysis and Visualization of data

Big data technologies are all about handling, processing and analysing huge amount of data. Apache hive is a data warehousing tool used build on Hadoop platform, which is used for querying and analysing the dataset using Structured Query like language named as Hive Query language. It analyses the data that has been stored in Hadoop file system.

Along with the predication, the analysis of data is going to be carried out using Apache Hive Query Language, it supports analytical processing functions like sorting, aggregating, filtering, joining and transforming of data. Here between the different attributes analytical processes will be carried out to provide meaningful result, and find the relevance between them along with the information which are not visible in dataset. Pre-processed and cleaned data has been used with most useful attributes to get inside of the data.

As for the visualization part, power BI is one of the most powerful tool for visualization of big data, It is easy to use as it provide different data connectors like spreadsheets, databases, cloud base platforms and so on. As the data being used is on movies, it contains the different information like title, genres of movies, cast, crew, companies name, language, budget, revenue and so on. The steps involve importing the cleaned data in .CSV or Excel format, creating visualization, filtration and report presentation.

3.5 Ethical Considerations

As the research and implementation has been carried out using the existing dataset and as this research is based more on content rather than users. There are no ethical considerations. But the dataset must have been created keeping in mind the ethical factors like privacy of user, consent from users and every one must have been provided with equal options and opportunities, Being fair to everyone.