**CAPSTONE PROJECT WORK REPORT**

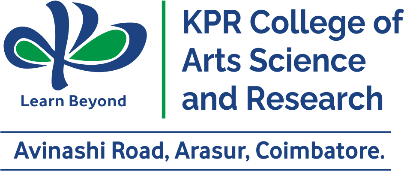
**Phase II**

**Movie Rating Analysis using Python**

Bonafide Work Done by

**ARUN S**

**REG. NO. 2028B0005**



Dissertation submitted in partial fulfillment of the requirements for the award of Bharathiar University, Coimbatore-46.

**Signature of the Guide**  **Signature of the HOD**

[ Mrs. Jayapriya P]

Submitted for the Viva-Voce Examination held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Internal Examiner External Examiner**

**CAPSTONE PROJECT WORK REPORT**

**Phase II**

**Movie rating analysis**

**ARUN S**

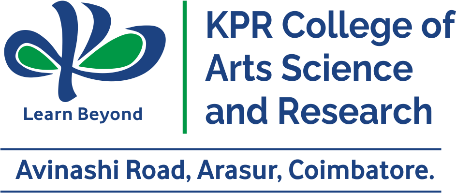
A report submitted in part fulfillment of the degree of

**B.Sc. in Computer Science with Data Analytics**

***Supervisor:* Mrs. P Jayapriya M.C.A,M.E.**,(**P hd**)

Associate Professor

Department of Computer Science with Data Analytics



**Department of Computer Science with Data Analytics**

**KPR College of Arts Science and Research**

(Affiliated to Bharathiar University, Coimbatore)

Avinashi Road, Arasur, Coimbatore – 641 407

|  |  |
| --- | --- |
| **TABLE OF CONTENTS** | **PAGE NO.** |
| ACKNOWLEDGEMENT | 4 |
| ORGANIZATION PROFILE | 5 |
| SYNOPSIS | 6 |
| **1. INTRODUCTION** | 7 |
| 1.1 Sentiment Analysis | 7 |
| 1.2 Functional requirements | 7 |
| 1.3 Movie success rate | 7 |
| 1.4 Alogrithm | 9 |
| **2. SYSTEM SPECIFICATION** | 9 |
| 2.1. Hardware Configuration | 9 |
| 2.3. Software Configuration | 10 |
| **3. SYSTEM STUDY** | 10 |
| 3.1 Data science work flow | 10 |
| 3.2 Source | 11 |
| 3.3 Transformation and analysis | 13 |
| 3.4 Visualaization | 13 |
| 3.5 Twitter only | 16 |
| **4. SYSTEM DESIGN** | 17 |
| 4.1 Sentiment Analysis on twitter data | 17 |
| 4.2 New features | 18 |
| 4.3 Collabrating on slack | 19 |
| 4.4 Challenges | 19 |
| 4.5 Newyork times API data | 20 |
| **5. CONCLUSION** | 25 |
| BIBLIOGRAPHY | 26 |
| APPENDIX | 30 |
| 1. source code 2. sample output | 24 |
|  |  |

## ACKNOWLEDGEMENT

In the accomplishment of completion of my Capstone Project Work Phase - Ion **Omicron Sentiment Analysis using Python** I would like to convey my special gratitude to **Dr. S. Balusamy, Principal of KPR College of Arts Science and Research** and **Mrs. P Jayapriya, Assistant Professor,Department of Computer Science with Data Analytics**. Your valuable guidance and suggestions helped me in phase - II of the completion of this project. I will always be thankful to you in this regard. I am ensuring that this project was finished by me and not copied.

**Student Signature**

**Place:**

**Date:**

### KPR COLLEGE OF ARTS SCIENCE AND RESEARCH



**(Affiliated to Bharathiar University, Coimbatore)**

**Avinashi Road, Arasur, Coimbatore – 641 407**

**ABOUT THE COLLEGE**

KPR College of Arts Science and Research is the latest addition to the KPR fleet. The College is located in a picturesque campus of about 11. Acres. The College is run by KPR charities under the leadership of our Chairman Dr. K.P. Ramasamy. The KPR Group is one of the largest industrial conglomerate in the country with interest in Textiles, Sugar, Wind Turbines, Automobiles and Education. The College was established in the year 2019 with a vision of providing top class education and life skills to students and thereby serve the nation and beyond. KPRCAS today offers 12 UG program in Management, Commerce and Computer Science streams. The Students of KPRCAS undergo intense training not only in the syllabus and curriculum of the affiliating University but are also trained in various areas. So that they emerge as industry ready graduates to meet the varying demands of the competing industries. Character building and Leadership qualities are inculcated into the students to make them responsible citizens focusing on the development of society and nation. A plethora of Clubs and Events encouraged the students to take part in sports and other cultural activities. KPRCAS offers three years undergraduate courses, which are exclusively for Business, Commerce and Computer Science Stream. The students are equipped with skills and knowledge needed to take up various leadership positions and to develop the society. Beyond Book Teaching help them to be professionals. KPRCAS emphasis on making the students academically brilliant, and also prepare them for the real corporate world. The learning curve begins here for the students of KPRCAS.

**ABOUT THE DEPARTMENT**

Bachelor of Computer Science with Data Analytics (B.Sc. (CS with DA)) was established in the year 2020. Data Analytics helps to raise the quality of data in the entire business system. The goal of data analytics is to construct the means for extracting business-focused insights from data This requires an understanding of how value and information flows in a business, and the ability to use that understanding to identify business opportunities. The primary aim of a data analyst is to increase efficiency and improve performance by discovering patterns in data. Data analysts exist at the intersection of information technology, statistics and business. They combine these fields in order to help businesses and organizations succeed. The students get exposed to Big Data, Business Intelligence, Data Mining, Data Visualization, Advanced Excel, Predictive Analytics and R Programming.

## SYNOPSIS

Twitter is a miniature writing for a blog site which gives phase to individuals to share as well as communicate their perspectives about point, activities, items plus other medicinal harms. Tweets can be arranged keen on assorted classes reliant on their significance through the tip looked. For genuine effecting of this structure python through NLP plus twitter informational compilation be used. In this project we are concerning feelings exploration in twitter tweet for omicron datasets to arrange the survey of all consumers whether it is positive, negative or impartial. The WHO designated variant of the coronavirus, B.1.1.529, as a variant of concern which has been named Omicron. Right after that, we saw an outbreak of tweets about the Omicron variant on Twitter. In this project, walk through the task of Omicron Sentiment Analysis using Python. Sentiment analysis is the task of natural language processing where we detect a positive, negative or impartial sentiment from a piece of text. Sentiment analysis is used by companies to analyze the opinions of customers about their products or services so that they can use the positive sentiments to market their products or services and the negative sentiments to improve the quality of their products or services, same strategy applied in our project.

## CHAPTER 1

## 1. INTRODUCTION

### 1.1 Sentiment Analysis

We usually come across movie rating websites where users are allowed to rate ad comment on movies online. These ratings are provided as input to the website rating system. The admin then checks reviews, critic’s ratings and displays an online rating for every movie. Here we propose an online system that automatically allows users to post reviews and stores them to rate movies based on user sentiments. The system now analyzes this data to check for user sentiments associated with each comment. Our system consists of a sentiment library designed for English as well as hindi sentiment analysis. The system breaks user comments to check for sentimental keywords and predicts user sentiment associated with it. Once the keywords are found it associates the comment with a sentiment rank. The system now gathers all comments for a particular movie and then calculates an average ranting to score it. This score is generated for every movie in the system. The system also sorts and displays top rating movies as per analysis and calculates a top ten list automatically. This provides an automated movie rating system based on sentiment analysis.

**1.2 Functional Requirements:**

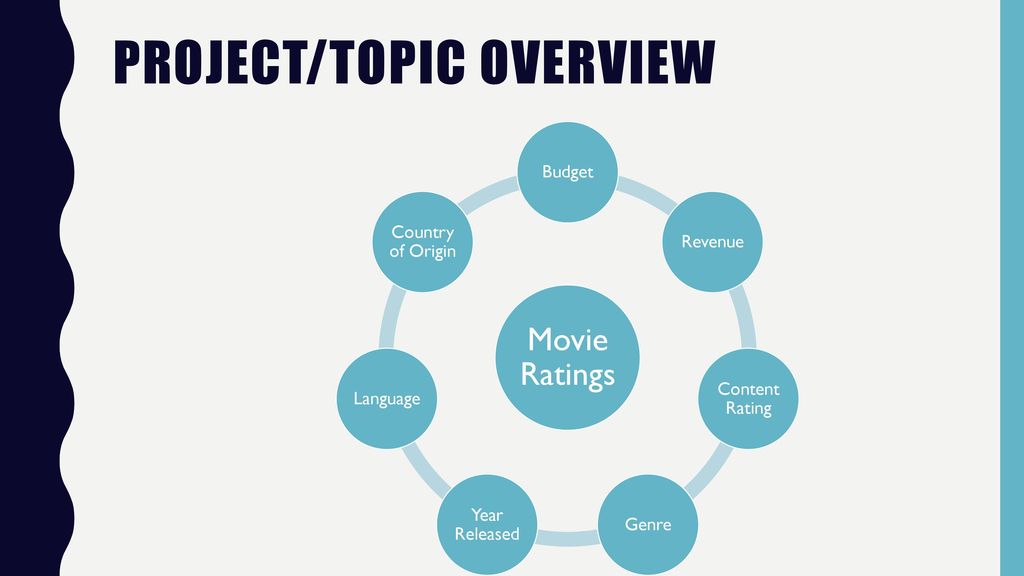
**FR1:**Users should be able to share his/her view about the stored movies provided in the system.

**FR2:**Users should be able to rate the stored movies provided in the system.

**FR3:**Users should be able to comment on stored movies only once. The system stores each comments of the users for further processing and find out the sentiments and their weightage and store it in database.

**FR4:**User can easily decide whether the stored movies provided by the admin/ system are good, bad or worst based on sentiment classification.

**FR5:**User can easily see/ analyze the better/top movies with the help of graphically representation of movie rating provided by the system.



**1.3 Movie Success Rate**

The film industry is one of the biggest contributors to the entertainment industry's unpredictability in success and failure. Because of quick digitization and the rise of internet-based life the film business is developing significantly as the average number of movies produced per year is

greater than 1000, therefore to make the movie profitable, it becomes a matter of concern that the movie succeeds. The success rate is the fraction or percentage of success among several attempts, and also, the average task success rate can be calculated either per participant or per task that

users complete correctly (Nielsen, 2006). Neural Networks have been extensively used in forecastingand prediction studies, it can, therefore, be employed for predicting the success and failure of the movies also (Sharma & Kaur, 2013). This study brings the understanding that the prediction of movie success is indeed possible with high percentages of accuracy, therefore, using a prediction engine, producers can

evaluate beforehand if the movie is worth investing in and accordingly make their decisions.

**1.4 Algorithm**

The designed algorithm is embedded in a developed web application that enables users (movie viewers) to write reviews for movies based on certain parameters. The web application was developed using Hyper Text Markup Language (HTML), JavaScript (JS), and the Python programming language. The administrator’s The reviews and ratings will serve as input to

the algorithm which will be used in predicting the

success of a movie.

The MR2P algorithm will further assign an average value to each review which will categorically state the user’s opinion on a particular movie, afterward, the average value of all reviews (A review) for a particular movie will be gotten as Alongside, the average rating of casts, genre, song, and the overall star rating will be gotten (A rating).Because the number of movie viewers increases.

## CHAPTER 2

## 2. SYSTEM SPECIFICATION

### 2.1. Hardware Configuration

|  |  |  |
| --- | --- | --- |
| **Operating System** | **Self-Hosted Technical Requirement** | **Cloud Technical Requirement** |
| Windows | Windows 8.1+ | Windows 8.1+ |
| Mac | Mac OS 10.14+ | Mac OS 10.14+ |
| Linux | Ubuntu LTS releases 18.04 or later | Ubuntu LTS releases 18.04 or later |
| **RAM** | 8 GB | |
| **HDD** | 1 TB | |
| **Processor** | 64-bit, four-core, 2.5 GHz minimum per core (If your dataset size is significantly larger than the medium dataset, we recommend 8 cores.) | |
| **Mouse** | Dell MS116 1000DPI USB Wired Optical Mouse | |
| **Keyboard** | Dell KB522 Business Keyboard-Black | |
| **Monitor** | Dell 24 Monitor-S2421HN in-Plane Switching (IPS) | |

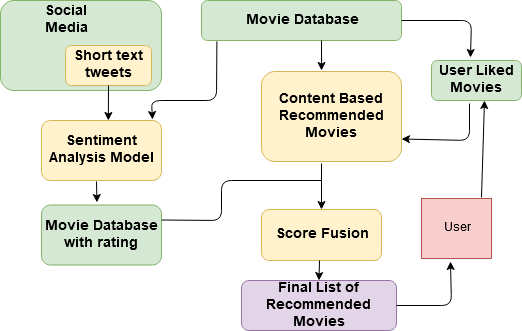
### 2.2. Software Configuration

|  |  |
| --- | --- |
| **IDE** | Anaconda |
| **Language Support** | Python 3.9 |
| **Platform** | Jupyter Notebook |
| **Browser** | Google Chrome Version 101.0.4951.67 |
| **Database** | MySQL 8.0.29 |

## CHAPTER 3

## 3. SYSTEM STUDY

# 3.1 Data Science Workflow

* We began by identifying our project goals and scope, assessing potential data sources, and retrieving the data.
* We first performed an API pull of a list of recent New York Times movie reviews.
* We then performed an API pull of recent tweets. We learned that Twitter provides only a limited search history for publicly available tweets. This resulted in an adjustment of our project scope.
* We then adjusted the date range of our New York Times movie list to contain only the movies with available Twitter data.
* We ran a sentiment analysis on the text of the tweets and the text of the movie reviews for the given set of movies.
* We created data frames for our data sources, cleaned and transformed the data, and performed analysis on the resulting datasets.

**3.2 1. Retrive list of movies review URLS ,other movies and attributes..**

# Sources

## Part 1. Twitter Search Setup

* We registered for a Twitter API key to perform the search of public tweets related to movies.
* This code chunk (Twitter Setup) was run first as it requires a PIN given by Twitter.
* We then ran the following chunks after entering the PIN for Twitter Authorization.

## Part 2. NYT Movie Review API

* We registered for a New York Times Movie API key to pull the movie review data.
* We used the opening date range 2016-04-21 through 2016-05-07.
* The resulting CSV is available here:

## Part 3. NYT Sentiment Analysis

For each URL pointing to a movie review, we then scraped the text of the review. Only the actual text of the review was considered. This was indicated by the class = ‘story-body-text story-content’ under the paragraph tag “< p >”. In addition, we retrieved the genre of the movie listed in the individual review. The genres are indicated by the following tag <span itemprop=‘genre’ class=’genre>. For movies categorized as more than one genre, we retrieved only the first genre tag.

We then ran the review text through sentiment analysis with indico, resulting in a 0-1 score, with 1 being most positive. For reproducibility of results, we read the review list and URLs from GitHub.

The resulting CSV is available h­­­­ere

## Part 4. Twitter Sentiment Analysis

Once we established the list of available New York Times movie reviews, we ran the Twitter search using the movie title as a search term. To get better Twitter search results for one-word movies, we added the word “movie” to our search term.

This chunk retrieved the Twitter search results for a term. searchTwitter has some other settings that could probably improve results.

We used the indico package to perform sentiment analysis, resulting in a score of 0-1 for each tweet, 1 being extremely positive.

We decided to use unique tweets only, therefore eliminating any retweet data.

For the benefit of our analysis, we filtered the Twitter search results to pull only the tweets which were coded with geographical location.

We chose the following cities at random and pulled tweets for only these cities.

|  |  |  |
| --- | --- | --- |
|  |  |  |

# 3.3 Transformation and Analysis

### New York Times Sentiment Analysis Litmus Test

We performed a sentiment analysis litmus test for New York Times reviews. The Times does not provide “stars” to rate their movie reviews but favorable reviews will be tagged with a “Critics’ Pick.” As a test for the sentiment analysis accuracy, we analyzed the results as follows:

We considered a “good” score any score >= 0.75 and expected the movie to be tagged with “Critics’s Pick” (value = 1).

We looked at Pick/Good (True Positive), Pick/Bad (False Negative), NotPick/Good (False Positive), NotPick/Bad (True Negative).

Percentage of False Positive: 35%  
Percentage of False Negative: 5%  
Percentage of True Positive: 15%  
Percentage of True Negative: 45%

The question when conducting sentiment analysis was our confidence in the results. Based on this quick analysis, we saw that the sentiment analysis score did not always match with whether the movie was recommended.

The number of false negative were indicators of how difficult it was to accurately measure the tonality and sentiment in a document. There were two movies that fell into the category of false negative. The movies were recommended by the New York Times critic; however, the sentiment analysis score was less than the threshold we considered. Upon reading the reviews, for these movies, “Viva” and “A Hologram for a King,” the sentiment analysis results were understandable. Both movies were dramas with elements of the review depicting difficult, dark, and negative terms.

The false positive results were more easily explained. The New YOrk Times gave neutral to positive reviews without recommending the movies. Also, a score of 0.75 may have been too low a limit to be expected to be a recommendation. We therefore considered scores below .50 as negative, scores between .50 and .80 as neutral, and above .80 as positive.

We reran the analysis based on these new limits. False negatives were scores < .50 but with a recommendation; false positives were scores >= 80 without a recommendation, and false neutrals were scores within interval (50-80) with a recommendation.

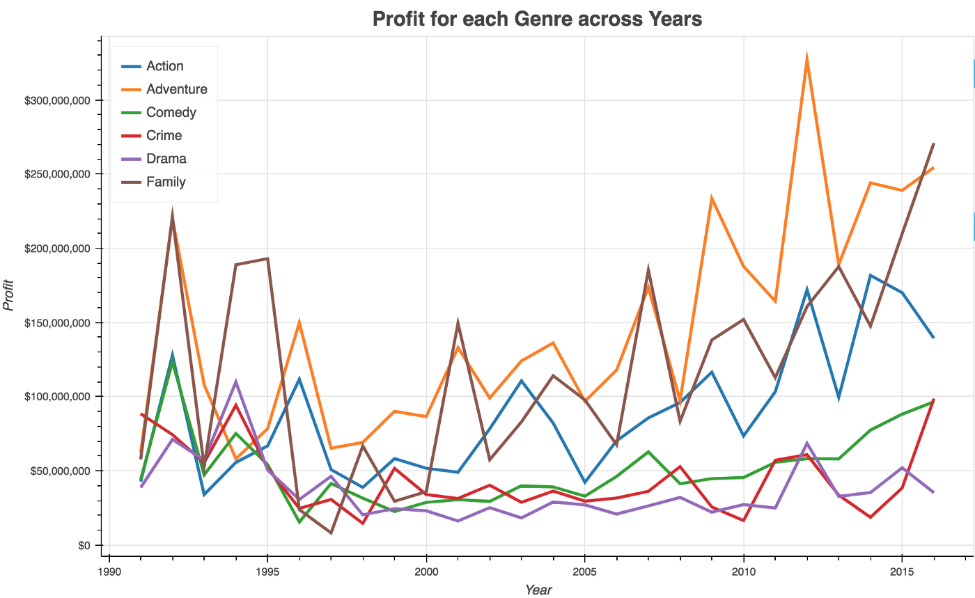
Percentage of False Positive: 27.5%  
Percentage of False Negative: 2.5%  
Percentage of False Neutral : 2.5%  
Percentage of True Positive : 15%  
Percentage of True Neutral : 12.5%

Percentage of True Negative : 40%

# 3.4 Visualization

## Box Plot

We compared New York Times review sentiment score and Twitter average sentiment score for the top 10 movies by tweet count.



We looked at average Twitter sentiment score by movie and by city.

### CHAPTER 4

### 4.1 Sentiment Analysis on Twitter Data

Performing sentiment analysis on Twitter data involves five steps:

1. Gather relevant Twitter data
2. Clean your data using pre-processing techniques
3. Create a sentiment analysis machine learning model
4. Analyze your Twitter data using your sentiment analysis model
5. Visualize the results of your Twitter sentiment analysis

# 4.2 New Features

We utilized the Twitter API to develop our collection of tweets. Using the twitter package we were able to pull tweets for locations by setting the geocode location and radius. OAuth was necessary because the uses the Developer API to search tweets based on our settings, it also required interaction by allowing the application access by putting in a key given by Twitter.

## Sentiment analysis machines

We tested a number of different sentiment analysis machines.

We started out trying to using the bag of words method to count positive and negative words within the tweet to get the positivity score. The problem with this method was that many of the words we had did not show up in tweets and the context of the tweet was typically misclassified.

We realized that we needed to get some help to get a more accurate sentiment analysis. We looked into several APIs such as IBM’s [Alchemy API](http://www.alchemyapi.com/api/sentiment-analysis). We ended up using Indico since it offered a free tier and they had developed an easy to use R package.

## Indico package

We used the Indico Sentiment Analysis API to analyze our text to return the positity on a scale from 0-1. We tested out several obviously bad and positive tweets and the response was very accurate.

Indico Benchmark: The API performs with 93% accuracy on the IMDB dataset (state of the art).

## Leaflet package

The Leaflet package was simple to use with our given list of locations with latitude and longitude. One drawback is the lack of a built-in zoom reset button. The popup option is a clean way to allow for label display without clutter.

## 4.3 Collaborating on Slack

We found Slack easy to use with a pleasant user interface and file sharing options. Slack may be more useful for larger groups. For our three-person group, we found that email was mostly manageable. Since we were already checking in on email, RPubs, and GitHub, Slack seemed like one more thing to check.

## Collaborating on GitHub

This was our first experience contributing to a repository using GitHub. It made sharing the code and version control of each change much easier than previous attempts. GitHub versioning control is powerful and can handle most merges seamlessly. However, if there is a conflict that GitHub cannot handle, the resolution must be handled through the Git command line and can be challenging.

# CHAPTER 5

# 5. CONCLUSION

Twitter has a limitation on interpretability. It is useful as a proof of concept but cannot be the sole source of information. Many tweets were commercial advertisements and had unhelpful links. Twitter is a useful tool to explore but we would not build a data product exclusively with it. Unless the search criteria can be narrowed to specific key words, we run the risk of pulling tweets that are not related to the indended search. The nature of tweets (shorter messages) makes it difficult to filter unwanted results. The Twitter data captured only 34 of our 40 New York Times reviewed movies. More obscure or independent movies, or movies with limited releases, were less likely to have mentions on Twitter. The New York Times does more in-depth cultural and artistic analysis than Twitter.

Lincoln, Nebraska had very few data points and skewed much of the results in that city. Also, a few movies had very few tweets and skewed the results of the movie score and we dropped in certain results. The results indicated that the New York Times did not reflect the tweet scores which we used as a proxy indicator of public opinion of the movies.

**BIOGRAPHY**

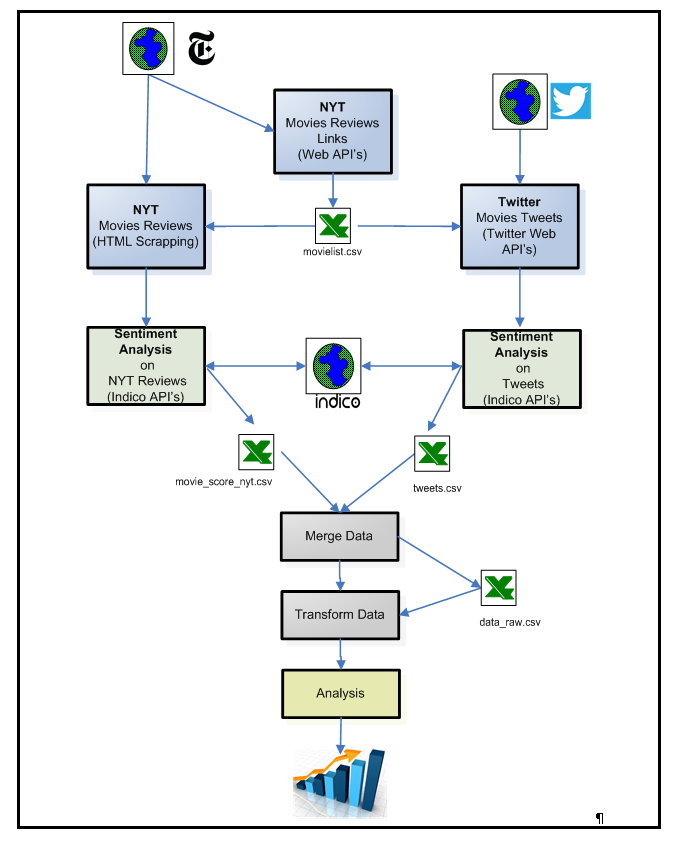
[1] Sampriti Sarkar, “Benefits of Sentiment Analysis for Businesses”, retrieved on: December 22, 2018, from: [www.analyticsinsight.net](http://www.analyticsinsight.net).

[2] ACME, “The Significance of a Film Review”, retrieved on: December 22, 2018, from: [www.revue-acme.com](http://www.revue-acme.com).

[3] Mshne, Gilad and Natalie Glance, (2006), “Predicting Movie Sales from Blogger Sentiment”, AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, PP 1-4.

# Appendix

**A ) data flow diagram:**



Source code:

import numpy as np

import pandas as pd

movies = pd.read\_csv("movies.dat", delimiter='::')

print(movies.head())

movies.columns = ["ID", "Title", "Genre"]

print(movies.head())

ratings = pd.read\_csv("ratings.dat", delimiter='::')

print(ratings.head())

ratings.columns = ["User", "ID", "Ratings", "Timestamp"]

print(ratings.head())

data = pd.merge(movies, ratings, on=["ID", "ID"])

print(data.head())

ratings = data["Ratings"].value\_counts()

numbers = ratings.index

quantity = ratings.values

import plotly.express as px

fig = px.pie(data, values=quantity, names=numbers)

fig.show()

data2 = data.query("Ratings == 10")

print(data2["Title"].value\_counts().head(10))

Sample output:

