Project I

# "Detecting Fake Reviews Using Opinion Mining"

Submitted in partial fulfillment of the requirements for the degree of

**Bachelor of Engineering**

by

**Shaikh Azamali Mohd Majid Laila Khatoon**     Roll No. 17DCO75

**Khan Khalid Ahmed Sajid Ahmed Shahida Nasreen**  Roll No. 16DCO54

**Ansari Mohd Faiz Mohd Siddique Shahjahan**     Roll No. 14CO17

Supervisor

**(Prof. Mukhtar Ansari)**



# Department of Computer Engineering,
**School of Engineering and Technology**
**Anjuman-I-Islam's Kalsekar Technical Campus**
Plot No. 2 3, Sector -16, Near Thana Naka, Khanda Gaon,
New Panvel, Navi Mumbai. 410206

**Academic Year : 2019-2020**

# CERTIFICATE



## Department of Computer Engineering,
### School of Engineering and Technology,
### Anjuman-I-Islam's Kalsekar Technical Campus
Khanda Gaon,New Panvel, Navi Mumbai. 410206

This is to certify that the project entitled *Detecting Fake Rating Using Opinion Mining* is a bonafide work of **Shaikh Azamali Mohd Majid Laila Khatoon, Khan Khalid Ahmed Sajid Ahmed Shahida Nasreen, Ansari Faiz Mohd Siddique Shahjahan** (Roll No. 17DCO75, 16DCO54, 14CO17) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineering** in **Department of Computer Engineering**.

**Prof. Mukhtar Ansari**                                    **Prof. Kalpana R. Bodke**

Supervisor                                                          Project Coordinator

**Prof. Tabrez Khan**                                          **Dr. Abdul Razak Honnutagi**

Head of Department                                                    Director

# Project I Approval for Bachelor of Engineering

This project entitled *Detecting Fake Rating Using Opinion Mining* by *Shaikh Azamali Mohd Majid Laila Khatoon, Khan Khalid Ahmed Sajid Ahmed Shahida Nasreen, Ansari Faiz Mohd Siddique Shahjahan* is approved for the degree of *Bachelor of Engineering in Department of Computer Engineering.*

Examiners

1. ..............................

2. ..............................

Supervisors

1. ............................

2. ............................

Chairman

............................

# Declaration

We declare that this written submission represents my ideas in my own words and where others ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

<div align="right">

Shaikh Azamali Mohd Majid Laila Khatoon

17DCO75

Khan Khalid Ahmed Sajid Ahmed Shahida Nasreen

16DCO54

Ansari Faiz Mohd Siddique Shahjahan

14CO17

</div>

# Abstract

**Tittle: Detecting Fake Reviews Using Opinion Mining**

In today's lifeline online marketing is on it's verge and keep increasing as people buy and sell products online. Digital marketing is an interesting and trending business platform. Sellers often post fake reviews on their products or pay people to post reviews and give higher rating, most consumer usually see and select products according to that product's rating and review which can be turn into dissatisfaction of consumer as he bought that product on the basis of fake reviews. To detect such reviews various methods are being used in past works. In this paper the method is being used is Sentiment Analysis (SA). SA has become one of the most interesting topics in text analysis, due to its promising commercial benefits. SA detects fake positive and fake negative reviews based on emotions in the opinion. In this study, we used machine learning algorithm Support Vector Machine (SVM) to detect those fake negative and fake positive reviews.

Reviews can be positive or negative which helps consumers to select product. This paper aims to classify movie reviews into groups of positive or negative polarity by using machine learning algorithms. For the movies data-sets we performed some data scrapping library like Beautiful-soup and Request to scrap movies data-sets and collected data-sets from websites which provides huge data-sets like dbmovies to perform our experiments. The measured results of our experiments shows that there is huge difference of 20%-30% between the actual rating of movies which is posted on movie reviewing websites and our experimental result.

Shaikh Azamali Mohd Majid Laila Khatoon
17DCO75
(Department of Computer Engineering)
(University of Mumbai.)

Khan Khalid Ahmed Sajid Ahmed Shahida Nasreen
16DCO54
(Department of Computer Engineering)
(University of Mumbai.)

Ansari Faiz Mohd Siddique Shahjahan
14CO17
(Department of Computer Engineering)
(University of Mumbai.)

# Contents

# List of Figures

# List of Tables

# Keywords And Glossary

**Keywords :** Fake Review Detection, Sentiment Analysis (SA), Machine Learning, Data Analysis, Deep learning, Data Scrapping, Data Visualization

## Glossary :

A:

Assumed - pretended, as in "an assumed name," referring to a pseudonym used to conceal one's identity (and, as a verb, to pretend); also, several unrelated meanings.

Algorithm - a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

API - a set of functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service.

Authentic - of undisputed origin and not a copy; genuine.

Analysis - detailed examination of the elements or structure of something.

Application - a software program that runs on your computer.


B:

Beguiling - deceptive, duplicitous, or diverting.

Bogus - not genuine.


C:

Collection - the action or process of collecting someone or something.

Contrived - false or unnatural.

Comparison - a consideration or estimate of the similarities or dissimilarities between two things or people.

Counterfeit - imitation or insincere.

D:

Delusory - deceptive.

Database - a structured set of data held in a computer, especially one that is accessible in various ways.

Data Mining - the practice of examining large pre-existing databases in order to generate new information.

Deep learning - part of a broader family of machine learning methods based on artificial neural networks. Learning can be supervised, semi-supervised or unsupervised.

Data - facts and statistics collected together for reference or analysis.

Detection - the action or process of identifying the presence of something concealed.

Dummy - imitation (and, as a noun, a mockup); also, several unrelated meanings.

E:

Ersatz - imitation or substitute (and generally inferior to the real thing).

Event - a thing that happens or takes place, especially one of importance.

F:

Factitious - artificial or false.

Fallacious - deceptive.

Feigned - fictitious, or not genuine or real.

Forged - imitation with intent to deceive; also, an unrelated meaning of being created by pressure and perhaps heat.

Feature - a distinctive attribute or aspect of something.

Fake - not genuine; imitation or counterfeit.

Fraudulent: deceitful.


G:

Genuine - actually having the reputed or apparent qualities or character.


I:

Integrate - combine (one thing) with another to form a whole.


J:

Jive - deceitful (said of talk); also, superficial or foolish, or other unrelated meanings.


L:

Log of events - a basic "log book" that is analyzed and monitored for higher level "network intelligence."


M:

Misleading - deceitful.

Monitoring - observe and check the progress or quality of (something) over a period of time; keep under systematic review.

Mobile Application - application software designed to run on a mobile device, such as a smartphone or tablet computer.

Mock - imitation, in the senses of not being genuine or real.

N:

Negative - consisting in or characterized by the absence rather than the presence of distinguishing features.

P:

Positive - consisting in or characterized by the presence rather than the absence of distinguishing features.

Put-on - pretended.

R:

Review - a formal assessment of something with the intention of instituting change if necessary.

Rating - a classification or ranking of someone or something based on a comparative assessment of their quality, standard, or performance.

S:

Sham - false, not genuine

Simulated - fake but intended or made to look genuine.

Specious - deceptively attractive, or appearing genuine or truthful.

Spurious - deceitful or imitation.

Synthetic - not genuine; also, several unrelated meanings.

V:

Vector - denoting a type of graphical representation using lines to construct the outlines of objects.

# Chapter 1

# Introduction

One of the very rapid growth area is Digital media marketing. Generally Digital media marketing provide facility for customers to write reviews related with its services. The existence of these reviews can be used as a source of information. For examples, movie makers can use it to make design decisions of their services but unfortunately, the importance of the review is misused by certain parties who tried to create fake reviews, both aimed at raising the popularity or to discredit the product. They share their thoughts on internet. Before watching anything, it is a normal human behavior to do a survey on that product. Based on reviews, customers can compare different movies and can finalize a product of their interest. These online reviews can change the opinion of a customer about the movie. If these reviews are true, then this can help the users to select proper product that satisfy their requirements. On the other hand, if the reviews are manipulated or not true then this can mislead user. This boosts us to develop a system which detect fake reviews for a product by using the text and rating property from a review. The honesty value and measure of a fake review will be measured by utilizing the data mining techniques. An algorithm could be used to track customer reviews, through mining topics and sentiment orientation from online customer reviews and will also blocked the fake reviews.

## 1.1 Problem Statement and Objective

### 1.1.1 Problem Statement

In recent years, online reviews have been playing an important role in decision making. This is because, these reviews can provide customers with large amounts of information about the goods or services. However, to promote fictitiously or lower the quality of the products or services, spammers may forge and produce fake reviews. Due to such behavior of the spammers, customers would be mislead and make wrong decisions. We are making this work to detect fake reviews of movies on various online websites. Thus detecting fake (spam) reviews is a significant problem. Opinion spamming refers to the use of excessive and illicit methods, such as creating a large volume of fake reviews, in order to generate biased positive or negative opinions for a target product or service with the intention of promoting or demoting it, respectively. The reviews created for this purpose are known as fake, spam or bogus reviews, and the authors responsible for composing such deceptive content are known as fake or spam reviewers.

### 1.1.2 Objective

The identified challenges motivate to bring up a solution to all the problems stated in the above problem statement section. Following are the objectives of the proposed approach and this thesis work:

- To detect fake positive, fake negative, real positive real negative reviews.
- To implement algorithm to get better Spam Detection i.e.; Account used, polarity of words using text blob, negativity and positivity of content, etc.
- Graphical representation of work.
- To presents Opinion Mining on Spam Filtered Data.
- To demonstrate the difference between the actual rating of the movie and the calculated resultant rating of the same.
- To present an algorithm that does Opinion Mining with Spam Detection.

## 1.2 Scope

our project is to build the system that can identify and distinguish between fake reviews and genuine reviews, our system will compare the actual rating of the movie and calculated resultant rating of particular movie. Our project will be working on machine learning algorithms that will do the good work of detecting reviews and identify reviews. In today's digital world the need of real and genuine opinion is needed on any product so it will save our time and money, and our project is doing the good work at it. As in digital media marketing to degrading or upgrading the value of movie uses of fake reviews are increasing and by seeing this problem we initiated our work and come up with this idea. Our system is build on top of machine learning and python libraries using algorithms our system works very accurately. The data-set we are importing is directly from IMDB website so the data are totally user based, unlike free data-sets available on internet this data-set is completely random and user's input so the evaluation will work on real data.

## 1.3 Technical Details

Technologies being used for the development of this application includes,

- Apache & Laravel 5.3 for web services.
- Vanilla JS, jQuery, AJAX for Client side validations .
- CSS3, Bootstrap, HTML5 for web designing.
- Flutter for mobile application development.
- MongoDB & Firebase Server for databases.
- BeautifulSoup & Requests libraries are used for data scrapping.
- Pandas are used for data storing in useful manner.

## 1.4 Innovativeness Usefulness

### 1.4.1 Movie Fake Reviews Detection.

Out system is about detecting movie's reviews as system like this isn't implemented yet, so this is the first of it's kind and can be useful to people who wants to save their time by reading about movies and make decision based on that.

### 1.4.2 Android Application for DFROM.

As our system is new we are making android application for this also. By making application, user don't have to go to website everytime. User can access and use our services in anytime on the go.

# Chapter 2

# Literature Review

## 2.1 Paper Title- Fake Review Detection using Classification.

In todays world, where Internet has become a household convenience, online reviews have become a critical tool for businesses to control their online reputation. Reviewing has changed the face of marketing in this new era. Nowadays, most companies invest money in mining the reviews to gain insights into customer preferences as well as to gain competitive intelligence and are hiring individuals to write fake reviews. The fraudsters activities mislead potential customers and organizations reshaping their businesses and prevent opinion-mining techniques from reaching accurate conclusions. Thus, it has become essential to detect fake reviews to bring to surface the true product opinion.

### 2.1.1 Weaknesses

- Only for e-commerce products eg: Amazon.
- architecture is complex.
- Computer aided generation of fake reviews are not detect.
- Naive Bayes and random Forest classification algorithm is applied which made the system slow.

### 2.1.2 How to Overcome

- To detect fake reviews of IMDB movie.
- We used SVM classification algorithm over Naive Bayes and random forest classification as this is faster comparatively.
- System is efficient and fast.

## 2.2 Paper Title- Detecting Fake Reviews through Sentimental Analysis Using Machine Learning Techniques.

This paper presents to classify movie reviews into groups of positive or negative polarity by using machine learning algorithms. This system analyse online movie reviews using SA methods in order to detect fake reviews. SA and text classification methods are applied to a data-set of movie reviews. More specifically, compare five supervised machine learning algorithms: Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN-IBK), KStar (K*) and Decision Tree (DT-J48).

### 2.2.1 Weaknesses

- Used five algorithm to detect the fake predictions which makes the system slow.
- Architecture is complex.

### 2.2.2 How to Overcome

- we used only one algorithm to detect the fake predictions which makes the system work rapidly and generate output.
- We also used the confusion matrix to count the real and fake predictions with known data.

## 2.3 Paper Title- An empirical study on detecting fake reviews using machine learning techniques.

This paper introduces a Reputation system in E-commerce (EC). The reputation systems aim at helping consumers in deciding whether to negotiate with a given party. Many factors negatively influence the sight of the customers and the vendors in terms of the reputation system. Moreover, the opinions obtained from users can be classified into positive or negative which can be used by a consumer to select a product.

### 2.3.1 Weaknesses

- Only used for E-commerce sites.
- four different algorithms to detect whether the given opinion is genuine or not.
- Using more algorithm to do same work require more amount of time.

### 2.3.2 How to Overcome

- Using a single but more accurate algorithm to do the task.
- Take less time to perform operation.
- Using SA on movies reviews to detect fake reviews and calculate precise result to give consumer satisfaction.

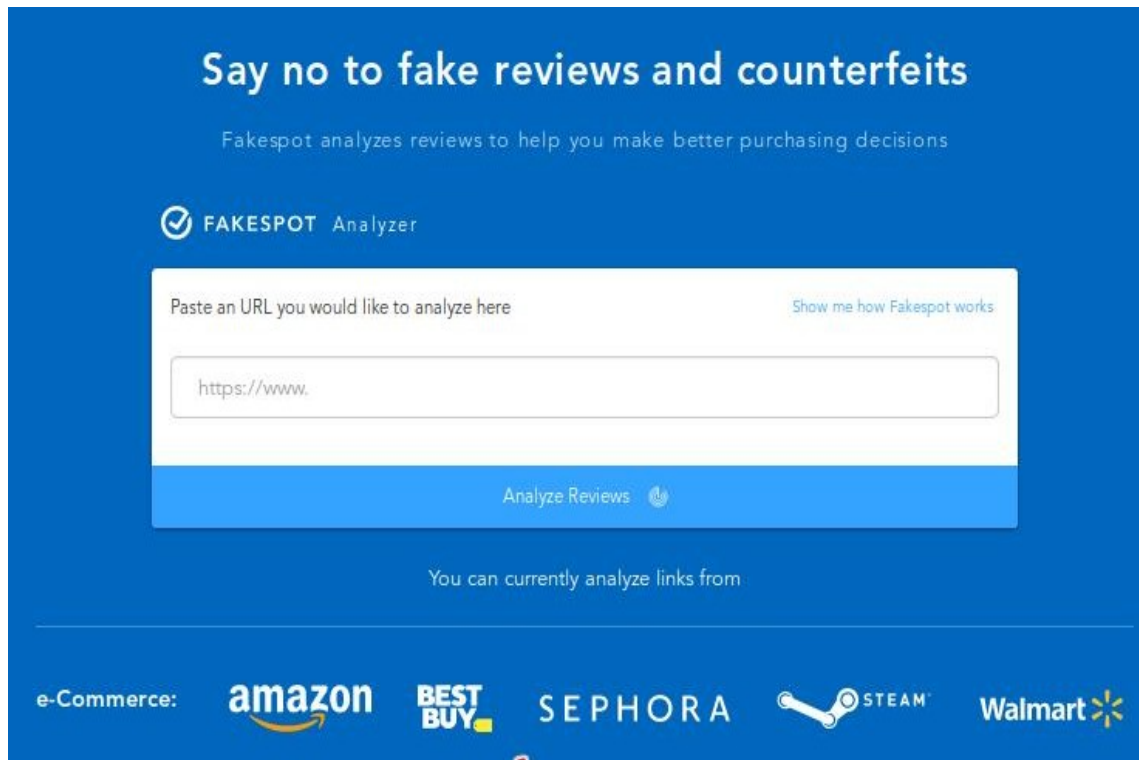## 2.4 Survey Existing System

### 2.4.1 Fakespot



Figure 2.1: Fakespot interface

- Fakespot works on products which are present on e-commerce websites e.g amazon.in, walmart.com, etc.

- It takes product by product's URL from browser.

- After taking product's information by fetching that from browser it works on it and gives the result.

- It shows differentiate between original rating of that product and calculated rating of that product.
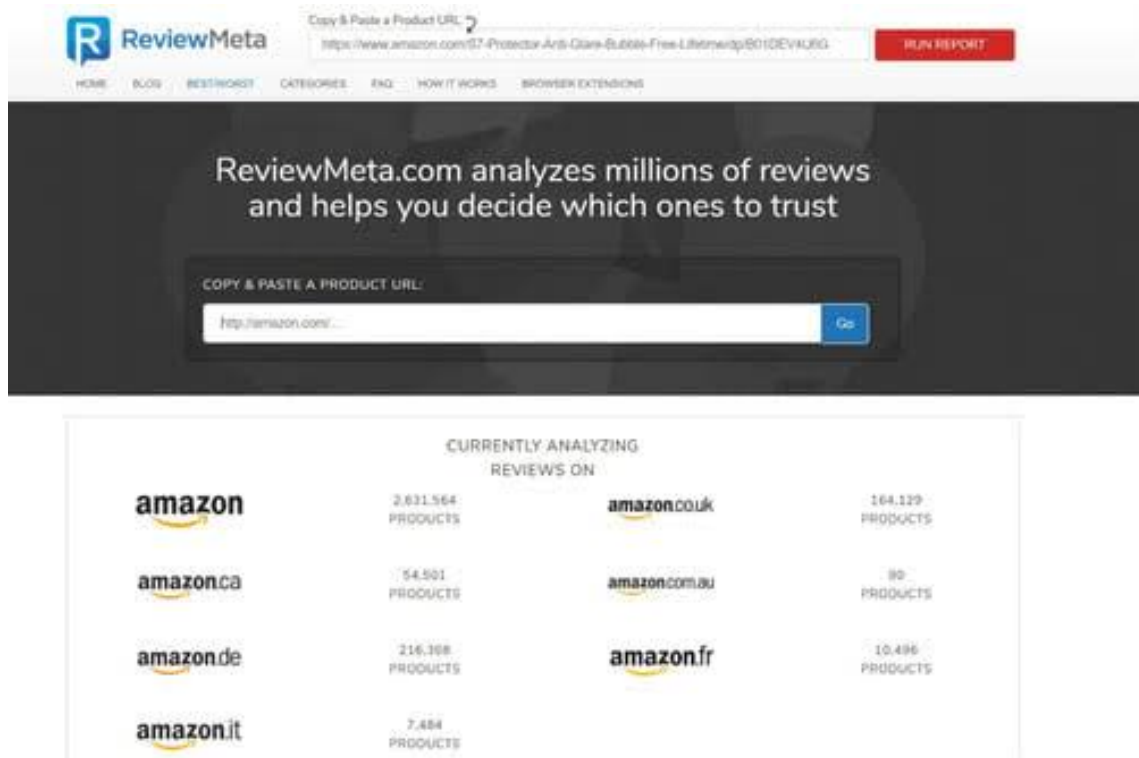
## 2.4.2 ReviewMeta



Figure 2.2: ReviewMeta interface

- This website only work with amazon products.

- It does not calculate/work on movie dataset.

- It shows the rating difference between the actual product rating and calculated rating of that product.

- It is faster than fakespot in some aspect as per our analysis.

### 2.4.3 The Review Index
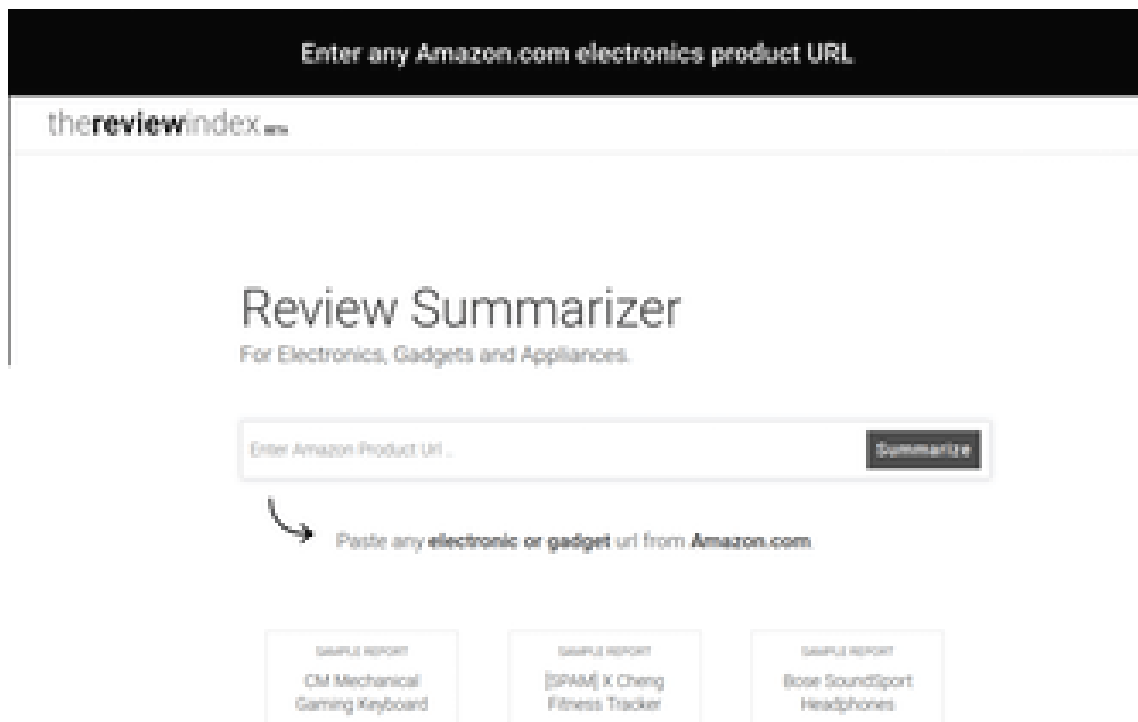


Figure 2.3: ReviewMeta interface

- The Review Index works on only Amazon and Flipkart products.

- It also does not work on movie dataset.

- It provids pluggin for google chrome and Mozilla Firefox to instant recognize the result of e-commerce product like amazon and flipkart.

- It displays the reviews and classify them and list them into spam and genuine review.

- UI/UX is easy and comfortable as compare to other.

## 2.5 Market Potential Competitive Advantage

| Features | DFROM App | Fake Spot | Review Meta | The Index Review |
|---|---|---|---|---|
| Detecting fake reviews of Movies. | ✓ | ✗ | ✗ | ✗ |
| Showing Spammer List. | ✓ | ✗ | ✗ | ✗ |
| Classification of Reviews eg: Good/Bad/Average. | ✓ | ✗ | ✗ | ✗ |
| Android Application. | ✓ | ✓ | ✓ | ✗ |

Table 2.1: Competitive Analysis

# Chapter 3

# Proposed System

## 3.1  Analysis

### 3.1.1  Framework

#### 3.1.1.1  Data Scrapping

A user search for movie information by giving a movie name in search box to get the movie rating of that movie.

User's requested movie name goes to the data scrapping module and then we apply our Data scrapping algorithm to fetch the user's desired movie information.

Now this raw movie information will be simplified according to our need and it will proceed on our next module which is Data pre-processing module
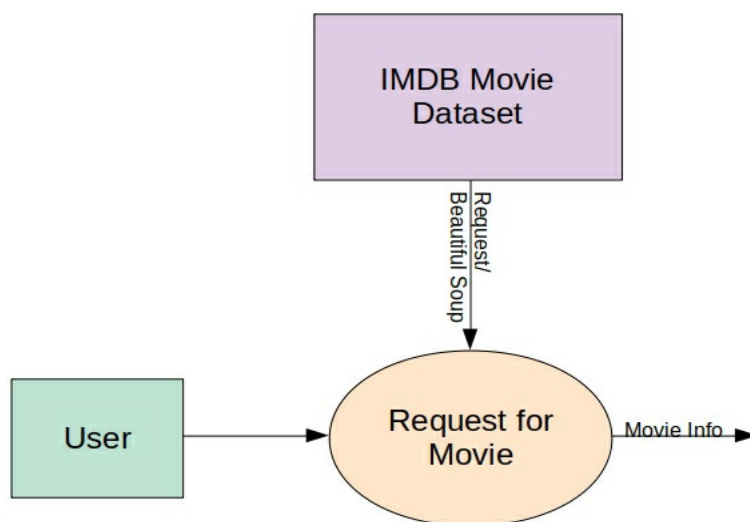


Figure 3.1: Data Scrapping Module

### 3.1.1.2  Data pre-processing

In Data pre-processing module we take the information of user search movie from the Data Scrapping module, on that information we apply algorithm for cleaning the data using various methods like word tokenizer, stop words removal and feature selection in this module.
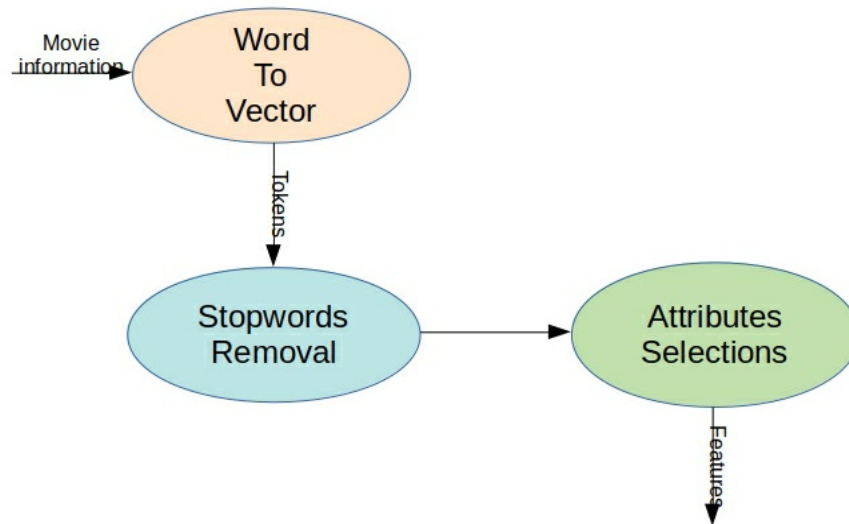
Figure 3.2: Data pre-processing Module

### 3.1.1.3  Comparison Result

After pre-processing the data we apply our algorithm to detect the fake reviews or rating in % and calculate the actual rating of the user search movie and show them our rating with IMDB rating.
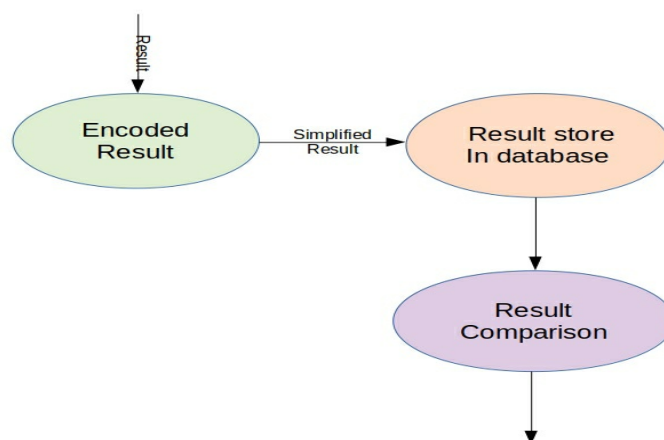
Figure 3.3: Comparison Result

### 3.1.2   Algorithm

The comparison in Table XI indicates that the classification accuracy of SVM algorithm was better than NB, KNN-IBK, K*, and DT-J48 algorithms.
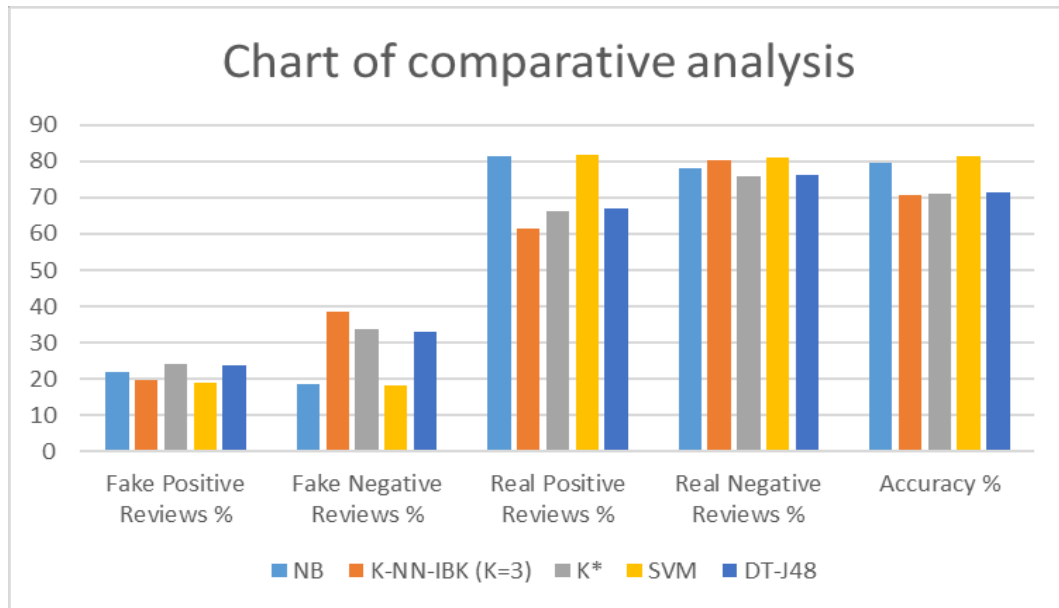


Figure 3.4: Comparative Analysis of Algorithm.

#### 3.1.2.1   CM : Confusion Matrix

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

The confusion matrix shows the ways in which your classification model is confused when it makes predictions.

It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

| | Class 1 Predicted | Class 2 Predicted |
|---|---|---|
| Class 1 Actual | TP | FN |
| Class 2 Actual | FP | TN |

Figure 3.5: Confusion Matrix.

**Definition of the Terms:**
Positive (P) : Observation is positive (for example: is an apple).
Negative (N) : Observation is not positive (for example: is not an apple).
True Positive (TP) : Observation is positive, and is predicted to be positive.
False Negative (FN) : Observation is positive, but is predicted negative.
True Negative (TN) : Observation is negative, and is predicted to be negative.
Falsse Positive (FP) : Observation is negative, but is predicted positive.

**Classification Rate/Accuracy:** Classification Rate or Accuracy is given by the relation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Figure 3.6: Confusion Matrix Equation.

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

#### 3.1.2.2   SVM : Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.
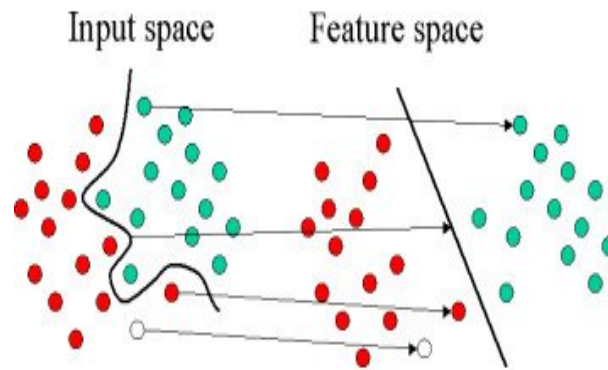


Figure 3.7: SVM example.

The process of rearranging the objects is known as mapping (transformation). Note that in this new setting, the mapped objects (right side of the schematic) is linearly separable and, thus, instead of constructing the complex curve (left schematic), all we have to do is to find an optimal line that can separate the GREEN and the RED objects.

### 3.1.3   Software Requirements

- Jupyter, Spyder

- Mongo-db

- Vs-code

- Android Studio

### 3.1.4   Hardware Requirements

- Laptop or Computer (Minimum 8GB RAM)

- Mobile (Minimum Android Versions : API 26)

- Internet Connection

## 3.2   Design Details
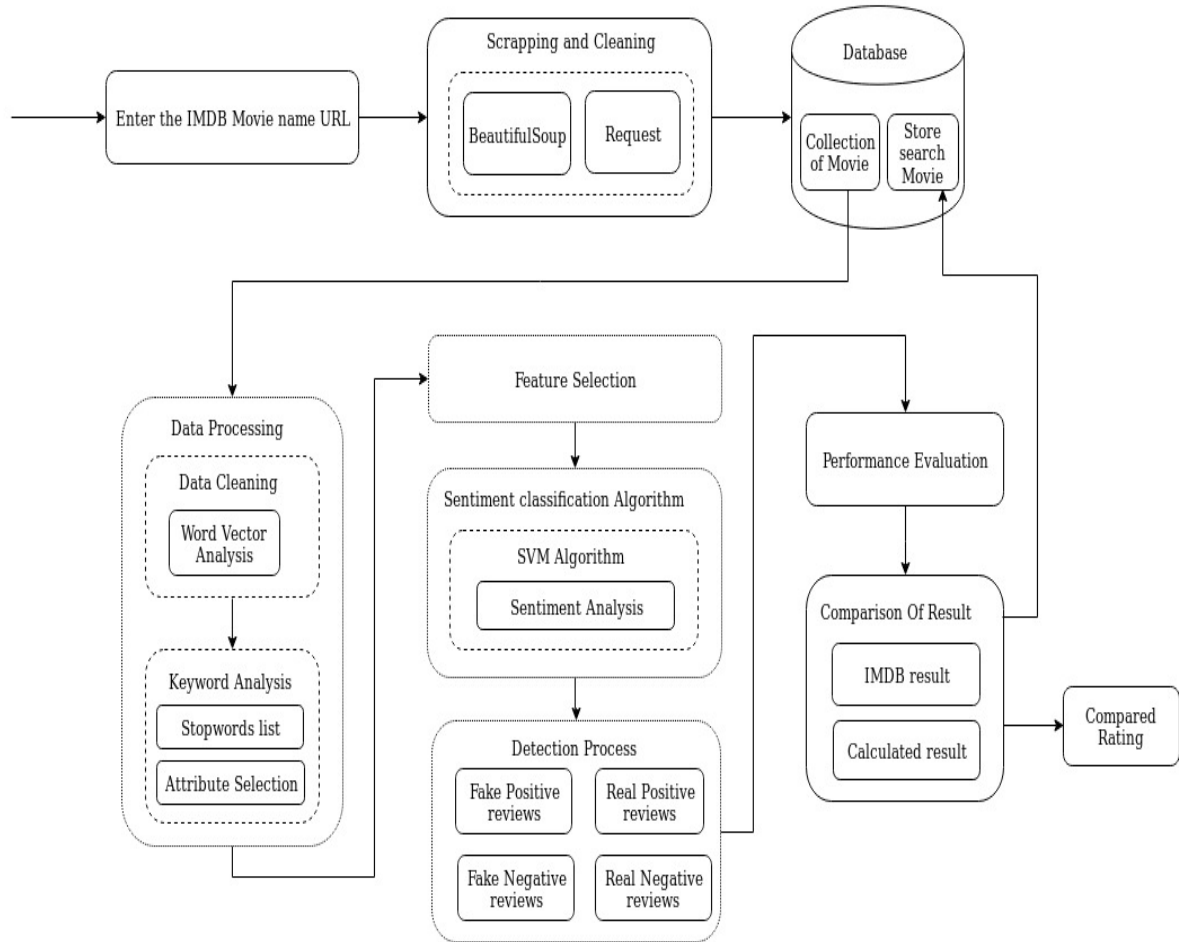
### 3.2.1   Project Architecture



Figure 3.8: System Architecture

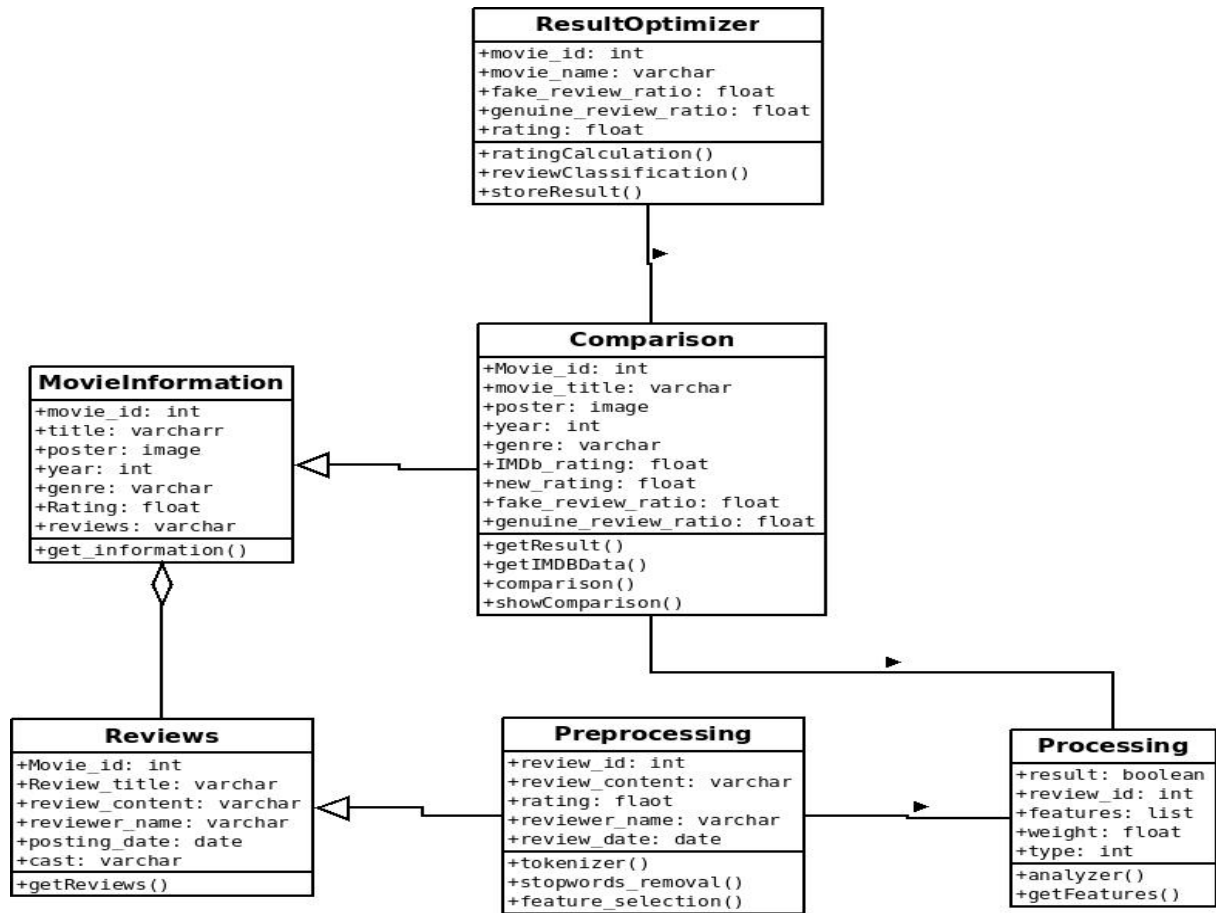# 3.3 Methodology

## 3.3.1 Class Diagram



Figure 3.9: Class Diagram for DFROM

Class diagram showing each modules interconnection and relation between how one module is interacting with others.
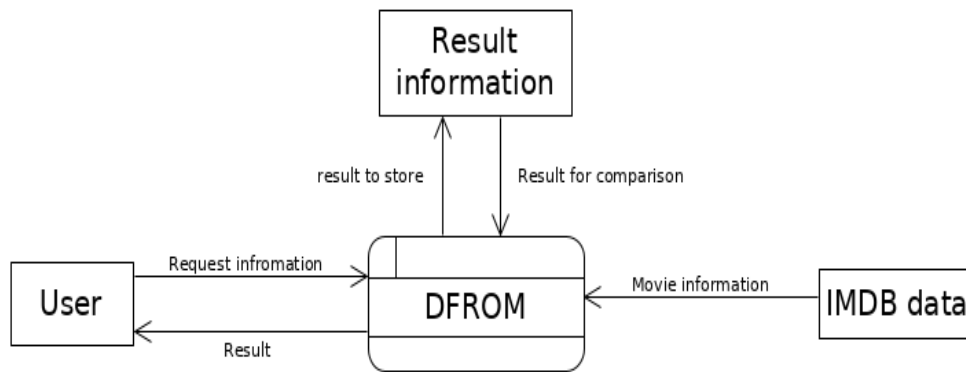
16

### 3.3.2 DFD Diagram



Figure 3.10: DFD Level 0 for DFROM

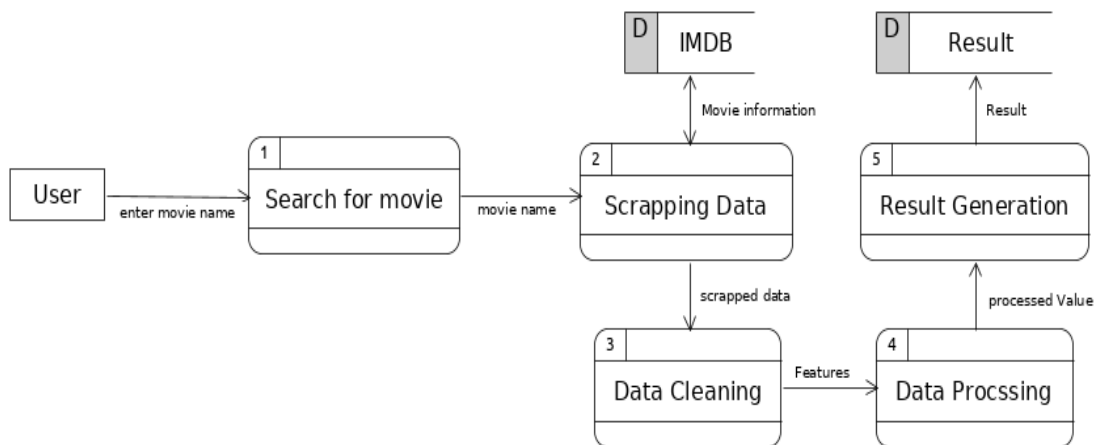DFD Level 0 depicting main outcome of the system, user can search the movie for comparing the rating from the IMDB.



Figure 3.11: DFD Level 1 for DFROM

In DFD Level 1 the user search the movie for getting correct rating, for that we scrap the data from IMDB and clean the data apply some algorithm to give the correct result to the user.
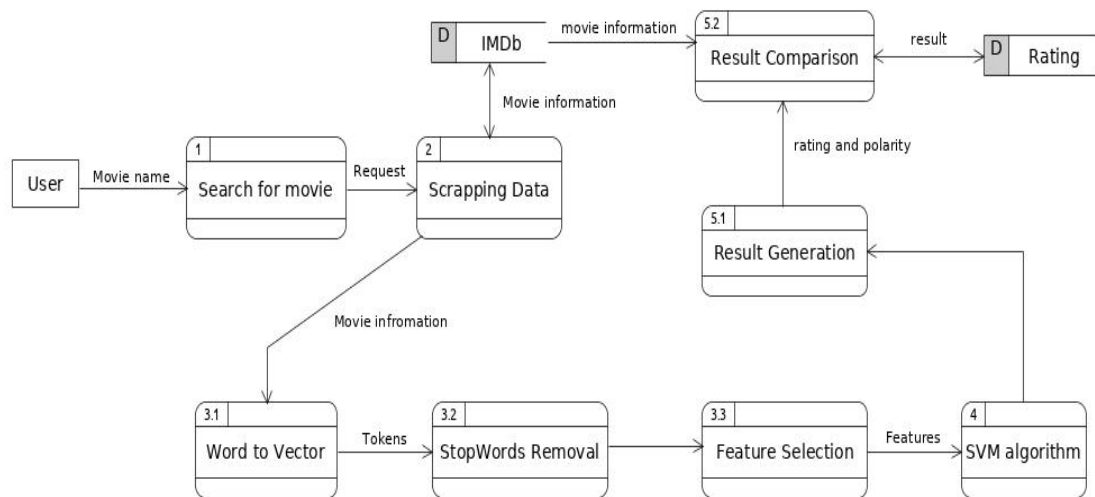
Figure 3.12: DFD Level 2 for DFROM

In DFD Level 3 the user search the movie and compare the rating from the IMDB. For that we use various algorithm like Sentiment Analysis (SA) to give the accurate result to the user .
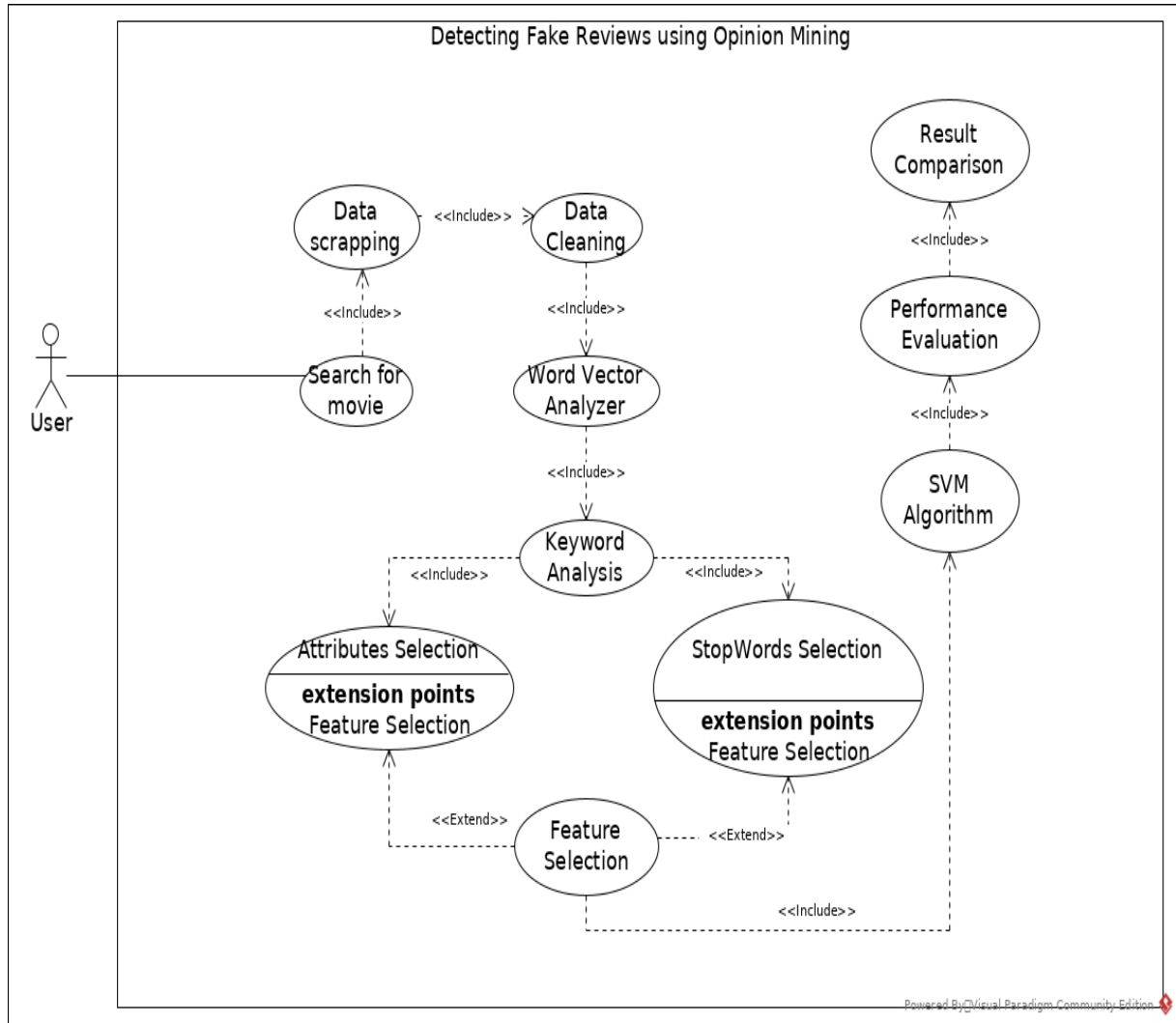
### 3.3.3 Use Case Diagram



Figure 3.13: Use Case Diagram for DFROM

Use case diagram depicting how user are interacting to the system. Also, how several processes are using each other functionalities via include & extends.
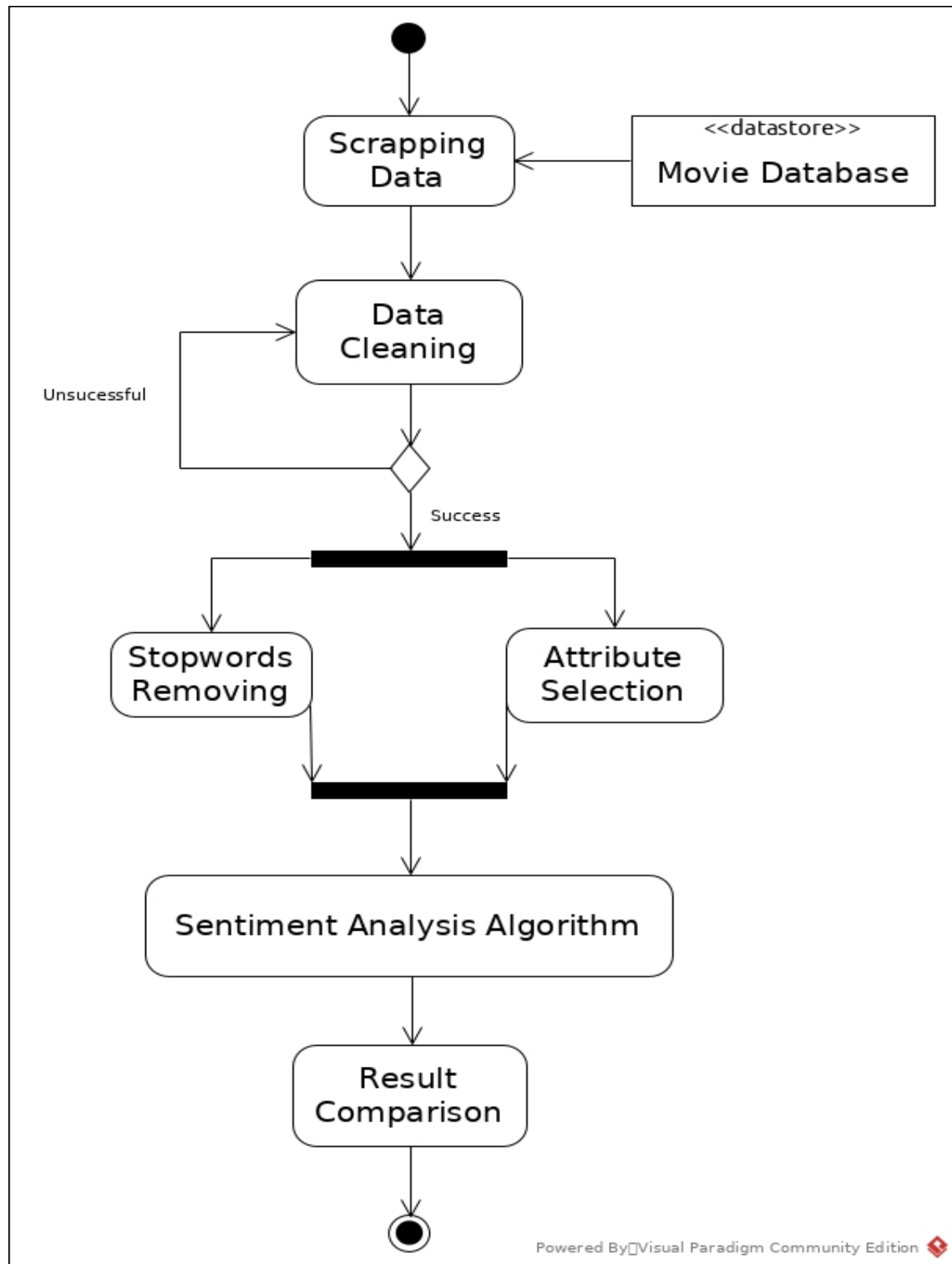
### 3.3.4 Activity Diagram



Figure 3.14: Activity Diagram for DFROM

Activity diagram, representing how the flow of a system is taking place beginning from search for movie. Process of data scrapping, Data cleaning, Feature Selection & applying Algorithm to appropriate rights & facilities.
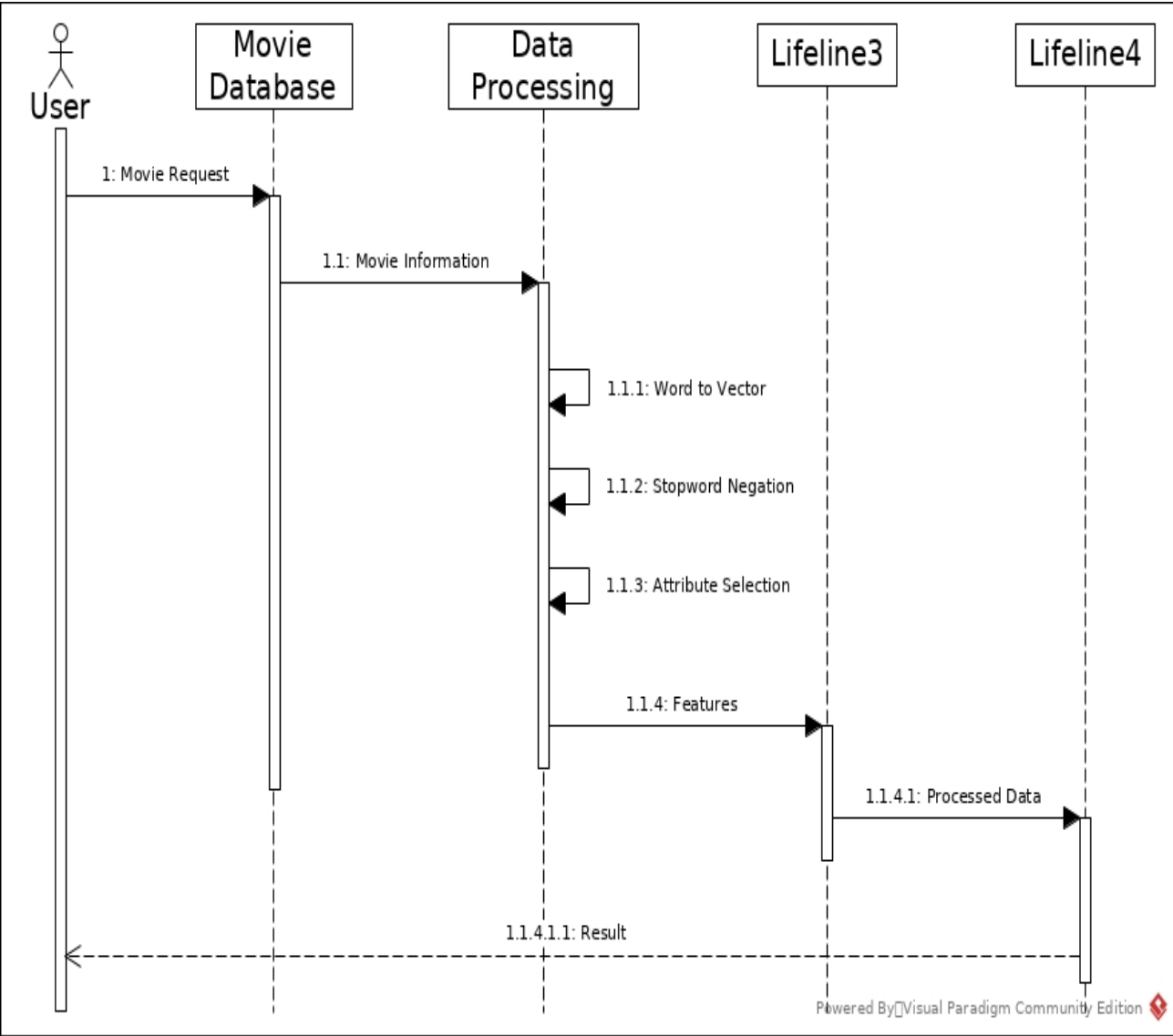
### 3.3.5 Sequence Diagram



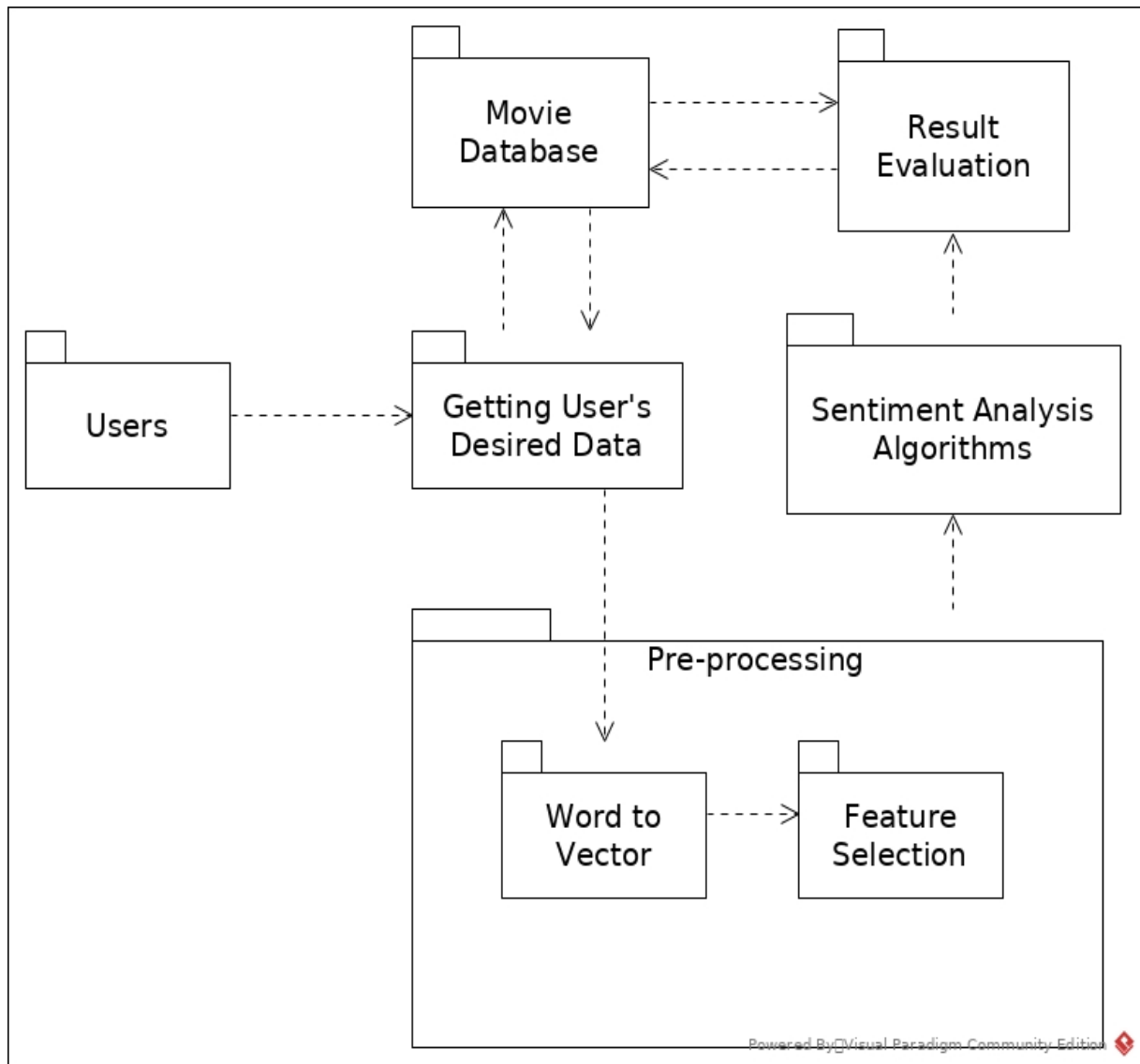Figure 3.15: Sequence diagram for DFROM

### 3.3.6 Modular Diagram



Figure 3.16: Modular diagram for DFROM

# Chapter 4

# Implementation Plan

## 4.1  Implementation Plan For Current Semester

### 4.1.1  Developing Key modules

- System Designing.
- Data scrapping using BeautifulSoup & requests libraries.
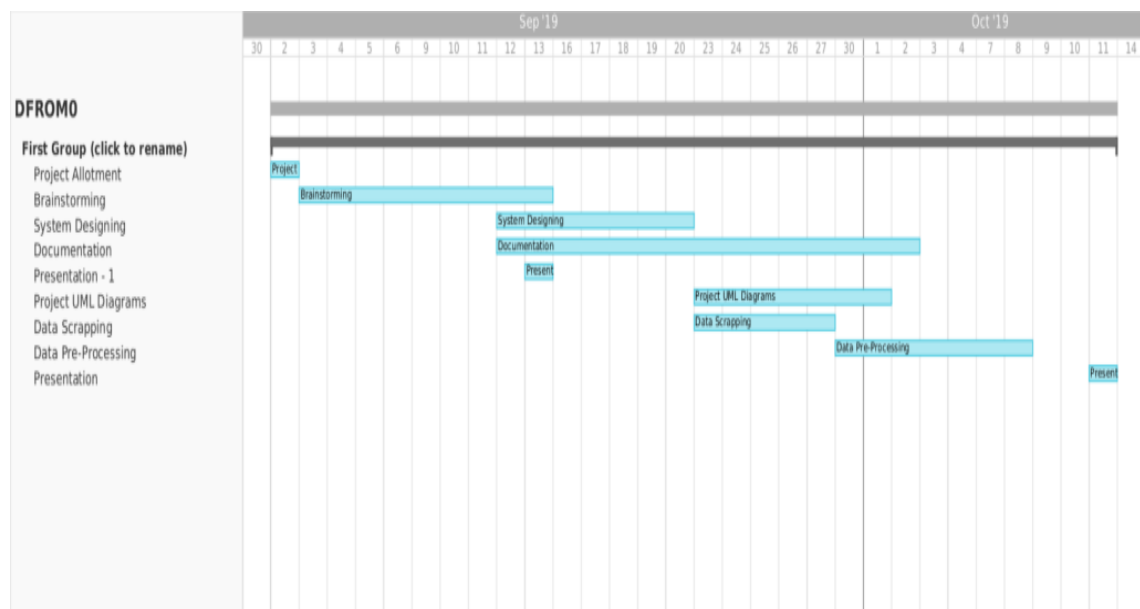- UML Diagrams.
- Data preprocessing.



Figure 4.1: Gantt Chart for the current sem

## 4.2 Implementation Plan For Next Semester

- Development of a full fledge system.
- Fake review Detection via Sentiment Analysis.
- Integration with mongodb database.
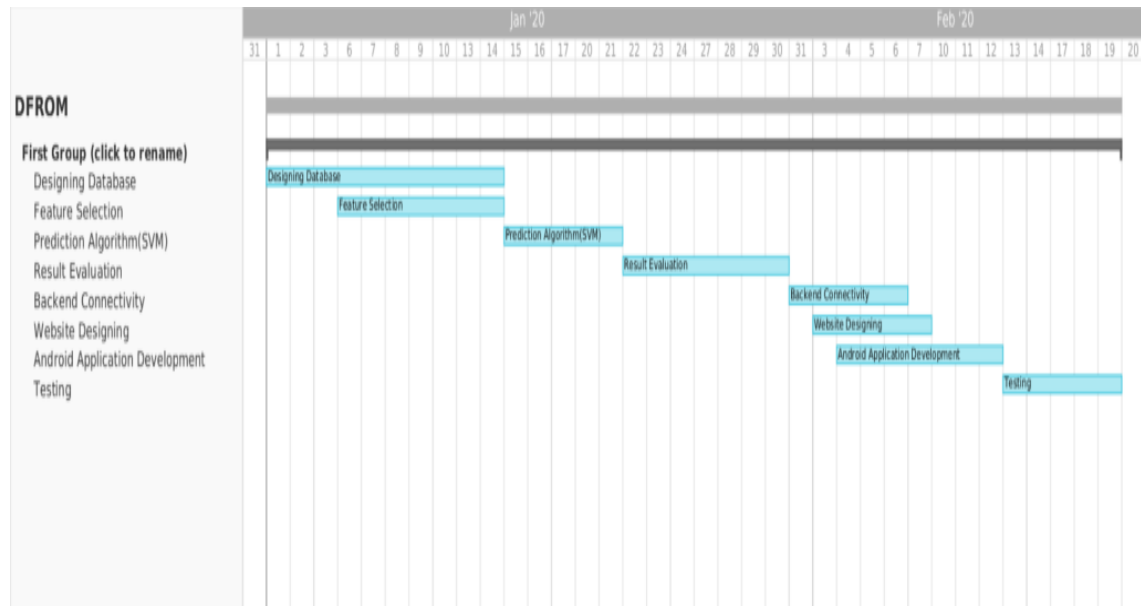- making a complete deployable application.



Figure 4.2: Gantt Chart for next sem

# Chapter 5

# Conclusion

Detecting Fake Reviews Using Opinion Mining (DFROM), is an movie review filtering system and it is cross platform digital solution to tackle the detecting fake reviews of movies.

We use the IMDB data to detect the fake reviews of the movie. As per the Previous researchers no one use the Original Data to detect the fake reviews all previous researchers use the automated generated data to detect the fake reviews.

We are using Sentiment Analysis to determine the movie reviews whether are fake or not. In sentiment Analysis we are using SVM (Support Vector Machine) Algorithm to get the accurate result.Previous researchers are used and implemented numbers of algorithms in their paper they have concluded that SVM is most promising among them. So we are using directly SVM based on their result it will apparently save processing time of our system.

# References

[1] *Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques*; Elshrif Elmurngi, Abdelouahed Gherbi, École de Technologie Supérieure, June 2018.

[2] *Fake Review Detection using Classification*; Neha S. Chowdhary, Anala A. Pandit, PhD, International Journal of Computer Applications, June 2018.

[3] *An empirical study on detecting fake reviews using machine learning techniques*; Elshrif Elmurngi, Abdelouahed Gherbi, École de Technologie Supérieure, August 2017.

# ACKNOWLEDGMENT

We would like to take the opportunity to express our sincere thanks to our guide **Prof. Mukhtar Ansari**, Assistant Professor, Department of Computer Engineering, AIKTC, School of Engineering, Panvel for his invaluable support and guidance throughout our project research work. Without his kind guidance & support this was not possible.

We are grateful to him/her for his timely feedback which helped us track and schedule the process effectively. His time, ideas and encouragement that he gave has helped us to complete our project efficiently.

We would also like to thank **Dr. Abdul Razak Honnutagi**, Director, AIKTC, Panvel, for his encouragement and for providing an outstanding academic environment, also for providing the adequate facilities.

We are thankful to **Prof. Tabrez Khan**, HOD, Department of Computer Engineering, AIKTC, School of Engineering, Panvel and all our B.E. teachers for providing advice and valuable guidance.

We also extend our sincere thanks to all the faculty members and the non-teaching staff and friends for their cooperation.

Last but not the least, We are thankful to all our family members whose constant support and encouragement in every aspect helped us to complete our project.

Shaikh Azamali Mohd Majid Laila Khatoon
17DCO75
(Department of Computer Engineering)
(University of Mumbai.)

Khan Khalid Ahmed Sajid Ahmed Shahida Nasreen
16DCO54
(Department of Computer Engineering)
(University of Mumbai.)

Ansari Faiz Mohd Siddique Shahjahan
14CO17
(Department of Computer Engineering)
(University of Mumbai.)