UGC & Govt. Approved
Sonargaon University (SU)
সোনারগাঁও ইউনিভার্সিটি (এসইউ)
WE WILL RISE UP WE WILL SHINE
147/1, Green Road, Tejgoan, Dhaka

Sentiment Analysis on IMDb Movie Reviews

Submitted by:

Gazi Fahim Hasan - CSE1601007009

Ferdush Bappy - CSE1601007015

Md. Jahirul Islam – CSE 1601007059

Md. Saiful Islam - CSE1601007040

Supervised by:

Arifur Rahaman

Lecturer

Department of Computer Science & Engineering

Sonargaon University

This project has been submitted to the Department of the Computer Science & Engineering at Sonargaon University in the partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering

April,2020

Sentiment Analysis on IMDb movie reviews

Department of Computer Science & Engineering

Sonargaon University

BONA FIDE CERTIFICATE

Certified that this project report "Sentiment Analysis on IMDb Movie Reviews" is the bona fide work of "Gazi Fahim Hasan - CSE1601007009, Ferdush Bappy - CSE1601007015, Md. Saiful Islam - CSE1601007040, Md. Jahirul Islam - CSE1601007059" who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Signature of the Supervisor

……..

Arifur Rahaman
Lecturer,
Department of Computer Science & Engineering

Signature of the Chairperson of the Department

……

Bulbul Islam
Associated Professor & head
Department of Computer Science & Engineering

ACKNOWLEDGEMENT

Firstly, we would like to all glorification and thanks to the almighty Allah for giving us trust to complete this project.

We would also like to express our sincere thanks to our Supervisor **Arifur Rahaman,** Lecturer, Department of Computer Science & Engineering, Sonargaon University for this constantly guidance, support and encouragement throughout the duration of this project, we consider it a privilege to have had the opportunity to work with him and share his valuable knowledge and experience. We are deeply grateful to him for detailed and constructive comments and for his important support throughout this work.

During this study, we have collaborated with many friends for whom we have greater regard, we wish to extend our warmest thanks to all those who have helped and supported ours. We express our sincerest thanks to the Department of Computer Science & Engineering, who gave us the opportunity to work with team and gave us untiring help during difficult moments.

We warmly thank to **Bulbul Ahmed,** Associated Professor & Head, Department of Computer Science & Engineering, Sonargaon University for his valuable advice and moral support, His expensive discussions around work and interesting explorations in operations have been very helpful for this study, We also express our thanks to Vice Chancellor **Professor Dr. Md. Abul Bashar,** who has given us the opportunity to study in such a decent environment.

Finally, we would like to dedicate this to our parents for their love, encouragement help throughout the project.

Sentiment Analysis on IMDb movie reviews

ABSTRACT

IMDb is the world's most popular and authoritative source of movie, TV and celebrity content, designed to help fans explore the world of movies and shows and decide what to watch.

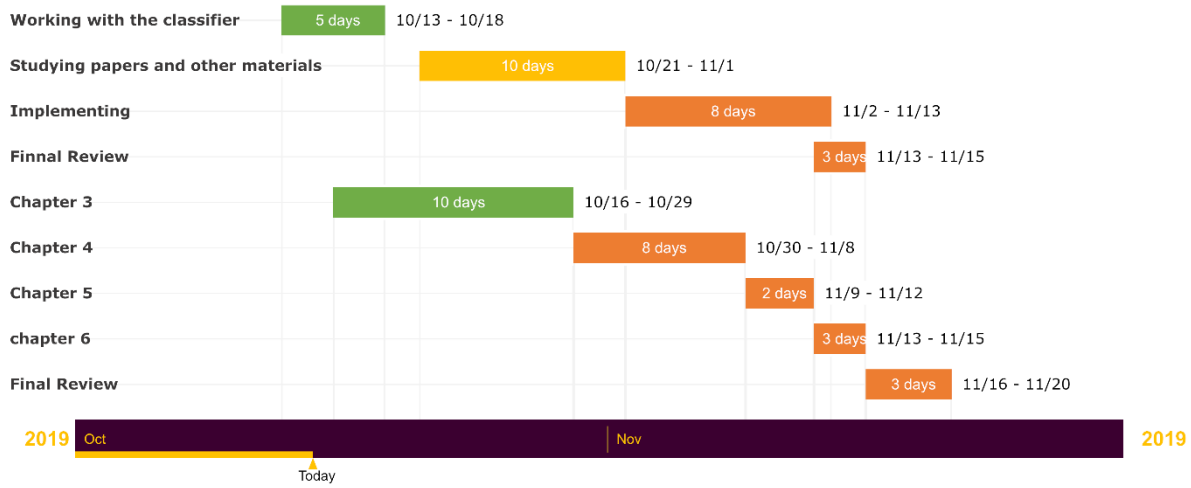
"IMDb Movie Reviews" has emerged as one of the online platforms to provide online movie reviews to the users. This is a project to implement a "Sentiment Analysis on IMDb movie reviews". To classify the reviews of every users either positive or negative. This project is basically analyzing of the user reviews to determine the user reactions on the basis of polarity and subjectively. The main purpose of "Sentiment Analysis on IMDb movie reviews" is to improve the relationship with the users and provide the better services to users as per users demand by analyzing the reviews of every users either positive or negative.

Sentiment Analysis also known as *Opinion Mining* is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Currently, sentiment analysis is a topic of great interest. Since publicly and privately available information over Internet is constantly growing, a large number of texts expressing opinions are available in review sites, forums, blogs, and social media. IMDb is an online database of information related to films, television programs etc. which also allows people to share and express their views .This project involves classification of IMDb movie reviews into two main sentiments: positive and negative. Naive Bayes algorithm will be used to develop a machine learning model which will predict the sentiment of the reviews.

The tools used to develop this "Sentiment Analysis on IMDb movie reviews" included Spyder, Python packages : NumPy, matplotlib, pandas, re, NLTK

Sentiment Analysis on IMDb movie reviews

Table of Contents



List of Figure

Chapter 01

Introduction

1.1 Motivation and Problem Statement

An article published in a newspaper or magazine that describes and evaluates a movie. Reviews are typically written by journalists giving their opinion of the movie. For many of us, reviews are like one written by our friends on facebook, are important in making our decision to watch a movie.

Similarly, these reviews are available to movie production companies which helps them- ¬To understand sentiment and check the popularity of their films ¬to figure out new marketing strategies and future directions. Human mind can read and understand whether a review is positive but for movie studios it is difficult to hire employees to simply read and judge movie opinions. So here comes Machine Learning to rescue - to process, reliably extract and classify the sentiment of unstructured movie reviews.

Regarding this scenario, sentiment analysis is a brilliant term of artificial intelligence for understanding the user or audience's sentiment in the basis of polarity and subjectivity. IMDb Movie has a lot of users and the reviews of a movie users are increasing day by day. "Sentiment Analysis on IMDb Movie Reviews" scraps the review data as data set and completes the classification of review texts by subjectivity or objectivity and negative or positive attitude of users. This project presents and empirical study of efficacy of classifying movie review. Thus this is a completely different approach by removing the unstructured data and then classifying reviews employing Natural Language Processing toolkit.

1.2 Project Scope

This section will describe the scope of the project, and briefly outline the preconditions and constraints that narrows down the focus area of this project.

Our first experiments, uses a corpus based on positive and negative movie reviews from IMDb. We gathered reviews from IMDb website and other available resources.

Our dataset contains 7500 reviews. It is intended as a larger benchmark dataset for sentiment classification, and consists of movie reviews, gathered from IMDb. The age of the datasets, does not concern us, since the theories tested in this thesis, can be

applied to newer datasets without any problems. The trustworthiness of the documents and data sources will not be a focus of this project.

Sentiment analysis methods till now have been used to detect the polarity in the thoughts and opinions of all the users that access social media. Researchers and Businesses are very interested to understand the thoughts of people and how they respond to everything happening around them. Movies Industries can use this project to evaluate their movies and services and to improve their movies and services. Sentiment analysis by this project has more possibility and accuracy rather than manual analysis of sentiment data virtually.

There is a lot of scope in analyzing the users review on the web. Now-a-days, there are lots of website which post the reviews of different movies provided by different reputed movies industries. Sentiment analysis will have to pace up with this change. "Sentiment Analysis on IMDb movie reviews" can help the movie industries to change strategies based on user feedback.

1.3 Project Objective

The objective of this project is to show how sentimental analysis can help improving the user experience over a system interface. The learning algorithm will learn what users' emotions are from statistical data then determine the sentiment. Reviews will be classified into positive and negative sentiments. "Sentiment Analysis on IMDb Movie Reviews" is the process of user review data mining, analyzing of users feedback and display a graph of the result which presents the polarity values basis of positivity and negativity.

1.4 Impact, Significant and Contribution

We've put impact, Significant and contribution in the center of our feedback cycle. The ability to understand positive feelings, or negative feelings has been difficult, for machines that lack feelings. Movie industries want to accommodate the sentiment analysis tools into areas of user feedback, marketing. Al though, by having this project, the industries will able to track the users and fulfill their parameter. The industries can easily detect the sentiment of the users by analyzing their review data.

1.5 Background Analysis

Sentiment analysis is a process to find opinions, emotions from text, reviews, tweets and other sources of natural language. All the opinion/emotion is captured using natural language processing. Now-a-days 'Natural Language Processing' is becoming more and more popular because of the amount data is getting bigger and bigger. Using natural language processing we can find our trends, popularity. This is big field for research on generating movie reviews of "IMDb Movie Reviews" from users sentiment collected from a web site whose posts the user reviews of "IMDb Movie Reviews". We have selected all the latest Reviews of users regarding movie as our experimental datasets.

The first commercial recommender system, called Tapestry introduces the term Collaborative Filtering. Tapestry was designed to recommend documents, gathered from newsgroups. The motivation for this product was to prevent users to be overwhelmed with documents. Later, the interest grew due to the relevance directly to e-commerce. Netflix, an online streaming video service, released a dataset containing 100 million ratings given by half a million users to thousands of movies. With this, they announced an open competition (Netflix Prize) for the best collaborative filtering algorithm in this domain, matrix factorization . The architecture of recommender systems and their evaluation on real-world problems is an active area of research . Applications are released in domains ranging from recommending webpages, music, movies, books and other consumer products the

users side, it is expected that serious e-commerce systems have some kind of a recommender system.

The goal of a Recommender System is to generate meaningful recommendations to a collection of users, for items or products that might interest them . Suggestions for movies to watch on Netflix or books to buy on Amazon, is real world examples of the results from recommender systems. How the systems works depends on the domain and the characteristics of the data available. For example, after watching a movie on Netflix you can rate the movie on a scale from (dislike) to (like). In addition, the system may have access to user-specific and item-specific profile attributes, like product description for instance. Recommender systems uses different methods and

approaches to analyze these data, to recommend the best product the user. Collaborative filtering systems analyze historical interactions alone, while Content-based Filtering systems are based on profile attributes. Hybrid approaches combine these two methods and matrix factorization uses the structures of a matrix to do recommendations. The next subsections, will describe these techniques in detail.In collaborative filtering systems, a user is recommended items based on the past ratings of all users collectively . User-based collaborative filtering works by collecting user feedback in the form of ratings for items in a given domain and exploiting similarities in rating behavior amongst several users. Finding a recommendation for an active user is done by choosing a subset of users based on their similarity with the active user. Thereafter a weighted combination of their ratings is used to produce predictions for the active user. The reviews are actually there personal opinions about a movie which are more informal and less technical I most cases. This system will take all the reviews of users from “IMDb Movie Reviews” and then process them using Natural Language Toolkit, do a sentiment analysis on these reviews to get the polarity and subjectivity which further leads us to understand of a review.

Chapter 02

Literature Review

2.1 Introduction

The literature review is an important part of a Project. It should be thorough and accurate. This project going to review on the article, book and internal resources to study about the "Sentiment Analysis on IMDb movie reviews".

2.1.1 Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. NLP is a component of artificial intelligence (AI). The development of NLP applications is challenging because computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands. Human speech, however, is not always precise -- it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

2.1.2 Natural Language Toolkit (NLTK)

The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). it provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language". It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. It also includes graphical demonstrations and sample data sets as well as accompanied by a cook book and a book which explains the principles behind the underlying language processing tasks that NLTK supports.

2.1.3 Stop Words

Stop words are generally thought to be a “single set of words”. It really can mean different things to different applications. For example, in some applications removing all stop words right from determiners (e.g. the, a, an) to prepositions (e.g. above, across, before) to some adjectives (e.g. good, nice) can be an appropriate stop word list. To some applications however, this can be detrimental. For instance, in sentiment analysis removing adjective terms such as ‘good’ and ‘nice’ as well as negations such as ‘not’ can throw algorithms off their tracks. In such cases, one can choose to use a minimal stop list consisting of just determiners or determiners with prepositions or just coordinating conjunctions depending on the needs of the application. Examples of minimal stop word lists that can use:

- **Determiners** –Determiners tend to mark nouns where a determiner usually will be followed by a noun.
  examples: the, a, an, another
- **Coordinating conjunctions** – Coordinating conjunctions connect words, phrases, and clauses.
  examples: for, an, nor, but, or, yet, so
- **Prepositions** – Prepositions express temporal or spatial relations.
  examples: in, under, towards, before

In some domain specific cases, such as clinical texts, we may want a whole different set of stop words. For example, terms like “mcg” “dr” and “patient” may have less discriminating power in building intelligent applications compared to terms such as ‘heart’ ‘failure’ and ‘diabetes’. In such cases, we can also construct domain specific stop words as opposed to using a published stop word list.

2.1.3.1 NLTK Stop Words

Natural Language Processing with Python Natural language processing (NLP) is a res earch field that presents many challenges such as natural language understanding. T ext may contain stop words like ‘the’, ‘is’, ‘are’. Stop words can be filtered from the te xt to be processed. There is no universal list of stop words in NLP research, however t he NLTK module contains a list of stop words.

2.1.4 Lemmatization

Lemmatization is an important aspect of natural language understanding (NLU) and natural language processing (NLP) and plays an important role in big data analytics and artificial intelligence (AI). Complex algorithms use the rules of linguistic morphology, in context with a particular language's vocabulary, to group words used in speech and writing by inflected forms. Deep learning is used to analyze and understand the grouping as a whole, so when any inflectional form of a word is mentioned, the base term's entire lemmatization is included.

In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatization algorithms is an open area of research. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set.The goal of lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

For instance:

am, are, is ⇒ be
car, cars, car's, cars' ⇒ car

The result of this mapping of text will be something like:
the boy's cars are different colors ⇒the boy car be different color

Chapter 03

System Implement Process

3.1 Process

The steps of our system model.

Those are:

- The first step is data scrapping from "....." using a script to scrape review as dataset.
- The second step is analyzing data to sort out the number of words, characters and other attributes to be needed for next steps.
- The third step is stop words count for natural language processing.
- The fourth step is data cleaning for reducing of unwanted data\symbols or special characters from the dataset.
- The fifth step is lemmatization of the words for grouping together the inflected from of a word.

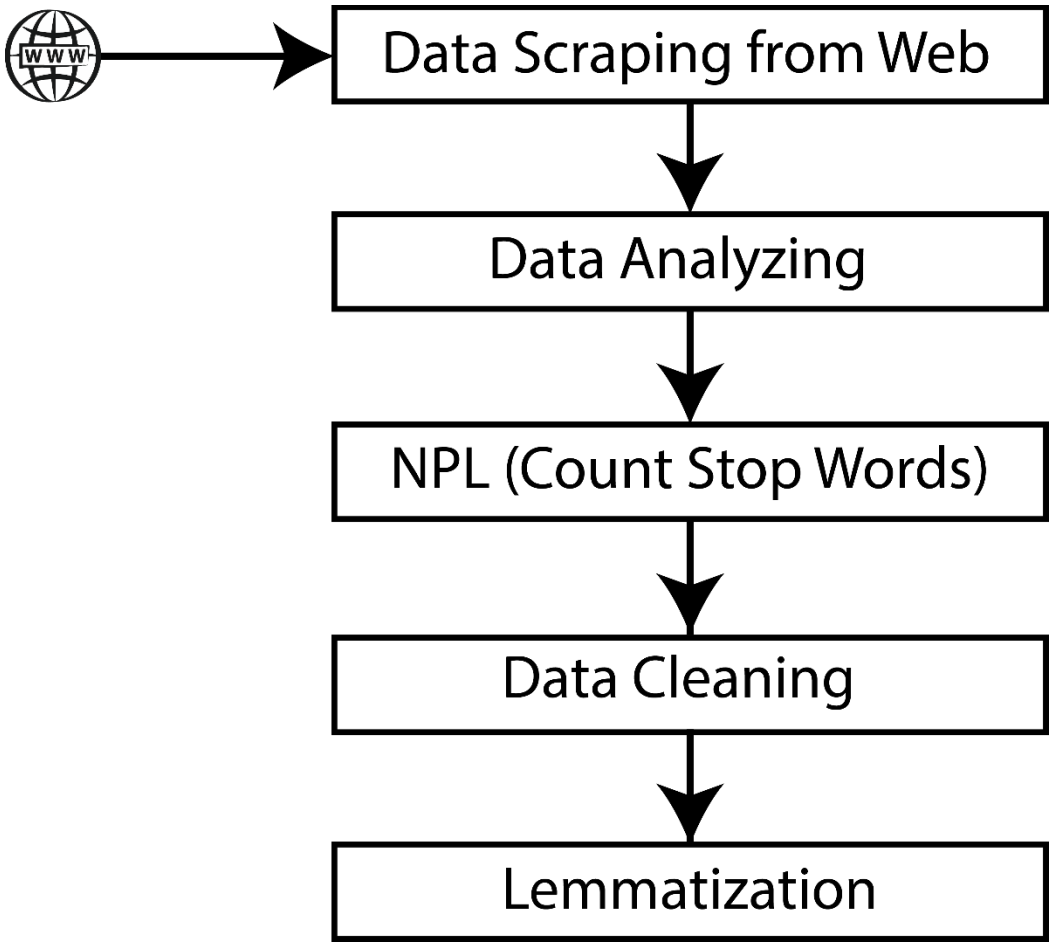


Figure-3.1: Process of system Implementation

.....

3.2 Tools

The main tools used to develop this project is python (OOP Language), Python library, package and module: NumPy, matplotlib, pandas, re and nltk.

Python: Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

NumPy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

Matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.

Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the

term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

RE (Regular Expression): Regular expression is a special sequence of characters that helps you match or find other strings or sets of strings, using a specialized syntax held in a pattern. Usually such patterns are used by string searching algorithms for "find" or "find and replace" operations on strings, or for input validation. Regular expressions are widely used in UNIX world. The Python module re provides full support for Perl-like regular expressions in Python.

NLTK: The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). it provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

3.3 Implementation Issues and Challenges:

Sentiment analysis classifies text as positive, negative or else objective, so it can be thought as text classification task. Text classification has many classes as there are many topics but sentiment analysis has only three classes. However, there are many factors that make sentiment analysis difficult compared to traditional text classification.

3.4 Timeline:

This project is estimated to be complete in a period of around 8 months. The following Grant Chart will show the timeline for each phases of the project. There are seven phases in this project and take different duration to complete. The initial planning will take around 30 days to complete, planning will take 10 days. After planning, the prototyping and design will carry out concurrently which prototyping is estimated to complete in 20 days and testing prototype is estimated to complete in 40 days. After testing prototype phase will be in implementation phase which take the longest days estimated as 80 days. When the implementation was done, the testing and debugging will carry out, each of it will take around 30 days.

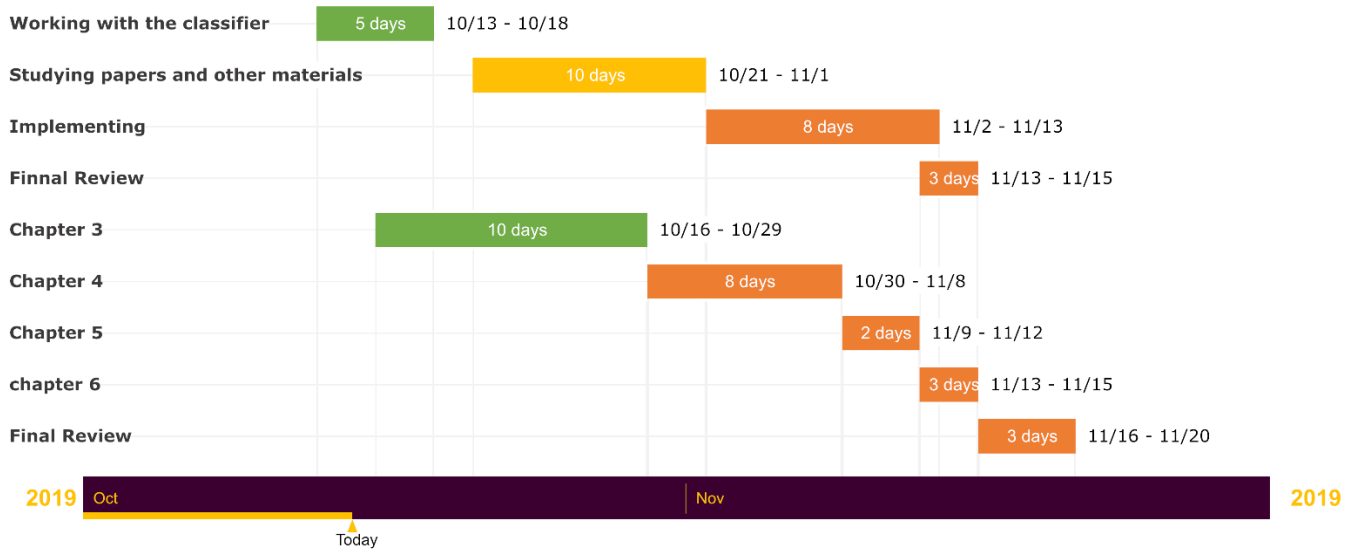


Table-3.4: Gantt chart of system Implementation

3.5 Requirement Specification

A software requirements specification (SRS) is a description of a software system to be developed. It is modeled after business requirements specification (CONOPS), also known as a stakeholder requirements specification (SRS). The software requirements specification lays out functional and non-functional requirements, and it may include a set of use cases that describe user interactions that the software must provide to the user for perfect interaction.

Data Collection: The raw IMDb dataset is structured in such a way that most of its attributes and information is organized and stored separately in compressed plain text files. For instance, all of the roughly 600,000 movie ratings from the database are stored in the compressed text file ratings. List (e.g. ratings.list.gz), which includes textual information about the data as well as a table of film rank, the number of votes and film titles. Thus, some sort of cleaning, integration and preprocessing is likely to be required in order to make good use of the data for the purpose of data mining through supervised machine learning techniques. The data was collected using IMDB (java movie database) which contains the IMDB movie dataset of more than 30,00000 movies in the dataset. The dataset was transferred to MySQL, in form of tables.

Sentiment Analysis on IMDb movie reviews

Data Analysis: The dataset is divided into training dataset and test dataset which contains the classes like Hit, Flop and Average and predicting variables like actor, actress, composer, genre, director producer and music director k-means clustering is used to analyze the training dataset to develop models which can be used for test dataset for analysis decision tree algorithm is used for predicting which factors.