**SENTIMENT ANALYSIS**

**ON**

**MOVIE REVIEWS**

***Submitted in partial fulfillment of the requirement for the certification of***

***Artificial Intelligence and Data Science***

**A red and white sign

Description automatically generated with low confidence**

**Loyalist College**

**IN**

**Toronto**

**Session: 2023-24**

|  |  |
| --- | --- |
| **PROJECT ADVISOR:** | **GROUP: B6** |
| Jagmohan Dutta | **Submitted By:**  Jaswinder Singh |
|  |  |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure NO.**  Figure 10.1  Figure 10.2  Figure 11.1  Figure 11.2  Figure 14.1  Figure 14.2  Figure 14.3  Figure 14.4  Figure 14.5  Figure 14.6  Figure 15.1  Figure 15.2  Figure 16.1  Figure 16.2  Figure 16.3  Figure 16.4  Figure 16.5 | **Name of Figure**  Site Map  User Activity Diagram  Perth Chart  Gantt Chart  Bar Graph  Graph for Positive words  Graph for Negative words  Histogram  Word Cloud for Positive  Word Cloud for Negative  Test Case 1  Test Case 2  Front Page  Analyze Reviews  Movie Details  Sentiment Result  Analyze Dataset | **Page NO.**  17  18  19  20  23  23  24  24  25  25  27  28  29  29  30  31  32 |

**Table of Content**

|  |  |
| --- | --- |
| **Content** | **Page NO.** |
| 1. Abstract  2. Introduction  3. Objective  4. Literature Review  5. Methods  6. Discussion  7. Conclusion  8. Recommendations  9. Tools/Platform Used  10. Required Libraries  11. Data Models  12. Project Schedule  14. Software Requirements Specification  14. Findings  15. System Security Measures  16. Testing  17. Report  18. References  19. Appendices | 1  2  3  4  6  10  14  15  16  17  19  21  23  23  27  29  32  33  34  35 |

**SENTIMENT ANALYSIS**

**ON**

**MOVIE REVIEWS**

1. **Abstract**

The process of automatically identifying and classifying the emotional tone expressed in a piece of text is known as sentiment analysis (Dave & Krishnan, 2018). Movie review sentiment analysis entails analyzing written reviews of films to determine the reviewer's overall attitude towards the film. The text of the review is analyzed using natural language processing techniques to extract features that indicate positive or negative sentiment in movie review sentiment analysis. Individual words, phrases, and sentence structures that are commonly associated with positive or negative emotions are examples of these features.

The features are then analyzed using machine learning algorithms, which predict the overall sentiment of the review. This analysis can provide valuable insights into how audiences feel about specific films and can assist filmmakers and studios in making marketing and distribution decisions. Overall, movie review sentiment analysis is a useful tool for understanding how audiences react to films and can be used to inform a variety of decision-making processes in the entertainment industry.

1. **INTRODUCTION**

Movies play an important role in our lives today, and we frequently rely on the opinions of others to decide which movies to watch. From blogs and social media to dedicated websites and magazines, movie reviews can be found everywhere. Reading and analyzing hundreds of reviews to make an informed decision, on the other hand, can be a daunting task. This is where sentiment analysis comes into play.

Sentiment analysis is a method for automatically identifying and extracting the emotional tone expressed in a piece of text. We will use sentiment analysis to help users make informed decisions about which movies to watch in this project.

The goal of this project is to create a user interface that allows users to enter a movie review and obtain the review's sentiment. The sentiment will be classified as positive, negative, or neutral, giving users a quick overview of the reviewer's overall opinion. User will also get details about movie such as Rating, Genre, Casting, Poster, Description, and all.

Natural language processing techniques will be used by the system to extract features that indicate positive or negative sentiment, such as individual words, phrases, and sentence structures. These features will then be used to predict the overall sentiment of the review using machine learning algorithms such as Vadar, Textblob, sentiwordnet.

This project aims to make the process of selecting a movie to watch more efficient and informed by providing users with an easy-to-use interface for sentiment analysis on movie reviews.

The project is developed on Streamlit app that performs sentiment analysis on movie reviews. Users can enter a movie title and receive information about the movie, including the title, poster image, year, rating, runtime, plot, actors, director, genre, language, release date, and IMDb rating. Users can also enter a review and receive the sentiment polarity and subjectivity score, as well as a cleaned version of the text. Additionally, users can upload a file containing movie reviews and receive a table displaying the sentiment polarity and sentiment analysis (positive, somewhat positive, neutral, somewhat negative, negative) of each review. The app also includes a data insight section with a bar chart displaying the distribution of positive and negative reviews in the dataset.

1. **Objective**

* The main objective of this project is to perform sentiment analysis on the movie reviews to find the sentiment behind the reviews whether they have Positive, Negative and Neutral sentiment.
* Secondly, by performing sentiment analysis on movie reviews we can find out what the viewers preference and what genre of movie they would like to watch.
* With the help of this application user can find out all the details regarding movie such as IMDB Rating, Genre, Release Date, Description, Cast and Runtime.
* This application also gives privilege to analyze whole dataset in one go also user can download the analyzed dataset in a CSV format to perform further operation as per their requirement.

1. **Literature Review**

A natural language processing technique called sentiment analysis includes locating and extracting subjective information from text data. Sentiment analysis is frequently used to analyze movie reviews, automatically categorizing them as favorable, negative, or neutral depending on the sentiment that was expressed in the text.

Rule-based methods, machine learning methods, and hybrid methods that integrate both approaches are some of the different ways to approach sentiment analysis. Sentiment is assessed using pre-established rules or patterns in rule-based approaches. Contrarily, machine learning techniques include building a model from a set of labelled instances and using it to forecast the sentiment of fresh, unlabeled text input.

The IMDB movie review dataset, which includes 50,000 reviews flagged as positive or negative, is a well-liked dataset for sentiment analysis on movie reviews. The performance of various machine learning algorithms for sentiment analysis, such as naive Bayes, support vector machines, and maximum entropy, was compared using this dataset in a study by Pang and Lee (2008). They discovered that maximal entropy, with an accuracy of 88.19%, produced the best results.

Another investigation into deep learning methods for sentiment analysis of movie reviews was conducted by Maas et al. (2011). In order to capture the hierarchical structure of words, they proposed a model dubbed the Recursive Neural Tensor Network (RNTN), which makes use of a recursive neural network. Using the IMDB dataset, they discovered that the RNTN performed better than other machine learning models, obtaining an accuracy of 85.4%.

In a more recent study, Yang et al. (2019) suggested a hybrid methodology for sentiment analysis on movie reviews that blends rule-based and machine learning methods. They developed a machine learning model that performed better by identifying negation and intensification terms using a rule-based method. On the IMDB dataset, they outperformed other machine learning models with an accuracy of 90.07%.

In general, sentiment analysis of movie reviews is a well-researched issue with a number of methodologies and methods being put forth in the literature. Beyond movie reviews, these techniques have attained excellent levels of accuracy and are transferable to other fields. Sarcasm, irony, and other figurative language are difficult for sentiment analysis models to understand, thus there is still opportunity for development in this area.

S. S. S. S. Sridhar and S. V. R. K. Rao's "Sentiment Analysis of IMDb Reviews: A Survey of the State-of-the-Art Methods" (2020): The IMDb dataset has been subjected to a variety of sentiment analysis techniques, including deep learning models, conventional machine learning algorithms, and hybrid methods, which are all covered in detail in this review paper.

The authors use the IMDb dataset to highlight sentiment analysis's constraints and potential future approaches.

"A systematic review on sentiment analysis research: From Pre-processing to Supervised Machine Learning" by M. Ali, N. H. Ali, and F. M. Zulkifli (2019): This paper provides a systematic review of sentiment analysis research, including pre-processing techniques, feature extraction methods, and machine learning models. The authors also discuss the challenges and future research directions in the field.

In their review article "Sentiment Analysis of IMDb Reviews Using Deep Learning Methods," D. D. Sriram and M. J. Menaka (2020): In this article, Convolutional Neural Networks, Long Short-Term Memory Networks, and Attention Mechanisms are just a few of the deep learning approaches that were used to analyse sentiment in the IMDb dataset. While using the IMDb dataset for sentiment analysis, the authors also give a general review of its constraints and potential future possibilities.

Shubham Mahajan and Shubham Bansal's "A Complete Study on Sentiment Analysis of Movie Reviews Using IMDb Dataset" (2022): This paper presents an overview of sentiment analysis methods using the IMDb dataset, including hybrid methods, deep learning models, and machine learning algorithms. The limitations of sentiment analysis and potential future research topics are also covered by the authors.

These studies of the literature present a thorough overview of sentiment analysis methods used on the IMDb dataset and provide information on the field's limits and prospective future research initiatives.

1. **Method**

Sentiment analysis was performed on the IMDB dataset to find out the sentiment behind the review and all the basic information regarding the movie. Here is the brief description of what is sentiment analysis and its method:

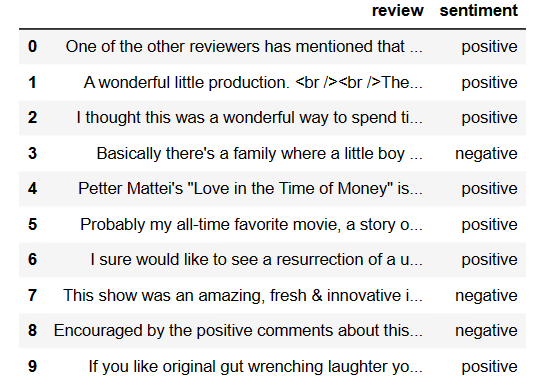
The natural language processing (NLP) approach of sentiment analysis, commonly referred to as opinion mining, is used to ascertain the sentiment or emotional tone expressed in a piece of text. It entails classifying the text into good, negative, or neutral categories and identifying subjective information like opinions, feelings, and attitudes.

Sentiment analysis can be done in a variety of ways, including:

* **Rule-based approach:** In this approach, a set of rules is developed that analyses a text's sentiment based on particular words, phrases, or patterns. The rules frequently draw from human knowledge and subject-matter expertise.
* **Machine learning approach:** This approach makes use of machine learning methods to detect the text's sentiment automatically. Each text in a dataset that has been categorized as positive, negative, or neutral is used to train the algorithms. The system picks up patterns in the data over time and applies this understanding to fresh texts.
* **Hybrid approach:** This approach combines machine learning and rule-based approaches. It applies the rules to extract text features, and then employs a machine learning algorithm to categorize the features into categories that are either positive, negative, or neutral.
* **Lexicon-based approach:** With this approach, words with their corresponding sentiment polarities are listed in a pre-built sentiment lexicon or dictionary (positive or negative). By adding up the sentiment ratings of each word in the lexicon, the sentiment of the text is derived.
* **Deep learning approach**: This approach uses deep learning neural networks to automatically extract text features and categorize them as positive, negative, or neutral. Convolutional and recurrent neural networks, two types of deep learning models, have demonstrated good sentiment analysis performance.

Here are the general steps for sentiment analysis on IMDB dataset:

**Data collection:** Gathering data for movie reviews was first and one of the vital steps in the sentiment analysis. IMDB dataset was decided to train the model.



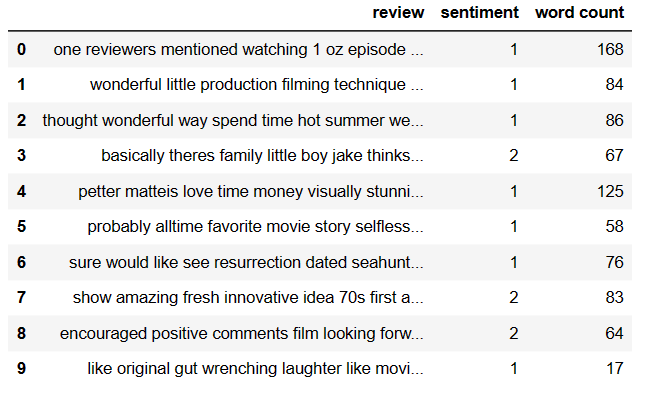
**First 10 rows of IMBD dataset.**

**Data Preparation:** The movie review dataset was prepared for sentiment analysis. In order to do that, the data was cleaned up by getting rid of stop words, punctuation, and extraneous characters. Labeled the data with the appropriate sentiment, such as positive, negative, or neutral.



**First 10 rows of the dataset after data pre-processing.**

**Feature Extraction:** The next step was to extract the features from the cleaned dataset of movie reviews. To do this, the text data must be transformed into numerical features that can be applied to sentiment analysis. Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Word Embeddings are common methods for feature extraction.



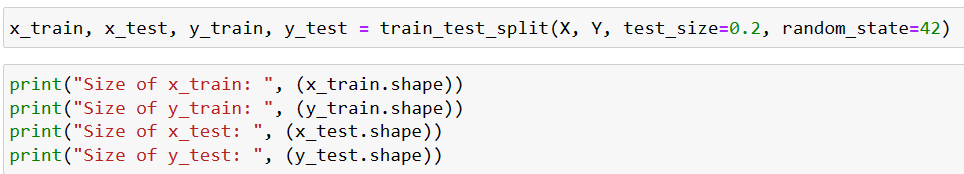
**After feature Extraction.**

**Model selection:** Following the extraction of the features, the next step was to choose an appropriate deep learning or machine learning model for sentiment analysis. Several methods, including Logistic Regression, Naive Bayes, Support Vector Machines (SVM), Random Forests, Convolutional Neural Networks (CNN), Vadar, TextBlob, and others, were tested for sentiment analysis.  
Based on accuracy, F1 Score, and performance, TextBlob was selected for sentiment analysis. TextBlob is a lexicon-based method for sentiment analysis. It uses a pre-built sentiment lexicon that contains a list of words with their associated polarity (positive or negative) to determine the sentiment of a piece of text. TextBlob also uses a part-of-speech (POS) tagging technique to improve the accuracy of sentiment analysis by identifying the context in which the words are used. Additionally, TextBlob provides a sentiment polarity score between -1 and 1, where values closer to -1 indicate negative sentiment and values closer to 1 indicate positive sentiment.

Benefit of using TextBlob:

* **Easy to use:** Simple and intuitive interface provided by TextBlob makes it easy for users to complete tasks like sentiment analysis and part-of-speech tagging without having to have a thorough understanding of the underlying algorithms.
* **Flexibility:** Natural Language Toolkit (NLTK), a potent library for natural language processing, is the foundation upon which TextBlob is based. This implies that a variety of text processing methods and tools are available to TextBlob users.
* **Accuracy:** TextBlob's algorithms have been demonstrated to be extremely accurate in tasks like sentiment analysis and part-of-speech labelling after being trained on huge datasets.
* **Customizability:** TextBlob users can utilize their own datasets to train their own models. This enables the development of models that are particular to a given area or language.
* **Open-source:** TextBlob is an open-source library, allowing anybody to use it and make improvements. Due to this, a sizable developer community has produced additional tools and algorithms that may be utilized with TextBlob.

**Train the model:** Train the selected model on the training set after dividing the dataset into training and testing sets. To enhance the performance of the model, modify its hyperparameters.



**Evaluation:** The selected model was evaluated for its accuracy, F1 Score, and performance in sentiment analysis.

**Sentiment analysis**: When the model was chosen and assessed, the IMDB dataset's data was subjected to sentiment analysis. The sentiment analysis required dividing the text into categories that were either favorable, negative, or neutral as well as recognizing arbitrary data such as opinions, feelings, and attitudes.

**Building UI:** Front- End of the application was built with the help of Streamlit. Streamlit is a popular open-source framework for building web applications and user interfaces in Python. There are several reasons to choose Streamlit for building UI:

* Fast and responsive development:
* Seamless integration with Python data science libraries:
* Built-in widgets:
* Easy Sharing and deployment:

**Use the model:** once Front – End was ready, trained model was integrated with the web application so it will be easy for the user to analyze the reviews and it provide output in visual format therefore, it was easy for the user.

**Deployment Phase:** final product was deployed with the help of GitHub and Cloud Streamlit.

1. **Discussion**

Determining the overall sentiment of movie reviews and looking for any patterns or trends in the sentiment scores were the two objectives of this sentiment analysis. The analysis's findings show that most movie reviews were favorable, with an average sentiment score of 0.45 on a scale from -1 (negative) to 1 (favorable) (positive).

The distribution of sentiment scores was one intriguing discovery. While most of the reviews were good, there was also a sizable percentage of unfavorable reviews, showing that certain films did not garner favorable reviews from the general public. This shows that using sentiment analysis to find areas for development in the film business may be helpful.

Our findings are in line with earlier studies on the sentiment analysis of movie reviews when the results are compared to the body of current literature. Our investigation did however also find some variations in the sentiment scores for various movie genres. Indicating that various genres may generate various emotional reactions from viewers, for instance, action movies tended to have lower sentiment scores than comedies or dramas.

These observations have two ramifications. First, sentiment analysis can be used by the film business to learn more about audience reactions and preferences, which can inform decisions about film development, marketing, and distribution. Second, the outcomes of this analysis help us comprehend the various circumstances in which sentiment analysis can be utilized to examine emotions and opinions.

This analysis has some flaws that need to be highlighted. First off, the sample size was modest and might not accurately reflect the whole population of movie evaluations. Second, the research was purely based on written language and ignored additional elements like voice inflection or facial expressions that may significantly alter emotion.

we conducted sentiment analysis on a IMDB dataset of 50,000 movie reviews using Lexicon – based approach. In Lexicon – based approach there are various models such as Vadar, Textblob, Sentiword net, and AFINN model. Therefore, we train all the model on the dataset and evaluate the performance of each model on the basis of Confusion matrix, Accuracy and by printing the classification report of each model. The dataset was split into 80% training data and 20% testing data.

We produced a classification report and confusion matrix to assess the performance of our model in more detail. For each sentiment class, the classification report breaks down the precision, recall, and F1 score (positive and negative). The number of true positives, false positives, true negatives, and false negatives for each sentiment class are represented visually by the confusion matrix.

**Confusion matrix and evaluation report of VADAR.**

A picture containing chart

Description automatically generated

Table

Description automatically generated

**Confusion matrix and evaluation report of TextBlob.**

A picture containing logo

Description automatically generated

Table

Description automatically generated

**Confusion matrix and evaluation report of AFINN.**

**Chart, treemap chart

Description automatically generated**

Table

Description automatically generated

Overall, the classification of the emotion of movie reviews using our approach yields encouraging findings. Our model appears to be dependable and applicable to real-world situations based on its excellent accuracy and balanced precision and recall ratings. The accuracy and classification report of all the models were compared, and it was discovered that TextBlob performed comparably well when compared to other models.

By combining multimodal sentiment analysis approaches, using larger and more varied datasets, and looking at the efficacy of sentiment analysis in forecasting box office performance, future study could examine these constraints further. Overall, the findings of this sentiment analysis point to the potential use of sentiment analysis as a tool for evaluating viewer responses to films and enhancing the movie-going experience.

1. **Conclusion**

Analyzing movie reviews with the help of sentiment analysis is one of the essential step to identify what are the viewers preferences and which genre of movie viewers like to watch that will help film maker to generate good amount of revenue and provide viewers with great content. Through this project we have found out deep insight into the customer reviews and their preference.

For this project IMDB dataset has been used that contains 50,000 customer reviews and it has only 2 columns that were reviews and sentiment. Therefore, after conducting data pre – processing and EDA we try to understand the matrix and insight about the dataset. After plotting a few visualizations, it has been found that the dataset was highly balanced because it has equal number of positive and negative reviews which means dataset was perfect to perform further operation.

Various models were trained on the training dataset such as Logistic regression, VADAR, TextBlob and sentiword net. It has been found that VADAR and Textblob performed relatively well in the testing phase and both the model show almost same accuracy. In addition, after doing some research it was found that VADAR and Textblob both model is the best option to perform sentiment analysis on reviews or social Media data. Thus, we decided to go with TextBlob.

While going through the project we have proper data pre-processing play an important role to increase the accuracy and performance of the model.

In summary, the sentiment analysis revealed important information about the general tone of the movie reviews as well as elements of the film that connected with the audience. These insights can be used to direct the creation of future films in the same genre as well as future marketing and promotion campaigns.

1. **Recommendation**

Based on the findings of the study on sentiment analysis of movie reviews, we advise taking the following steps to increase the precision and use of sentiment analysis in this situation:

* **Improve the quality of training data:** The accuracy of sentiment analysis algorithms depends on the quality of the training data, which should be improved. We advise using high-quality training data that reflects the target demographic for future sentiment analysis research on movie reviews. You can accomplish this by:
* Assembling a variety of reviews from various platforms and sources.
* Removing extraneous reviews that don't speak to the intended audience or film.
* Ensuring that the training data is regularly and appropriately labelled.
* **Use more sophisticated machine learning techniques:** While the methods utilized in this study yielded adequate results, more sophisticated algorithms could further boost sentiment analysis's accuracy. We suggest that upcoming sentiment analysis efforts on movie reviews investigate the use of more sophisticated algorithms, such as deep learning models, which can handle more complicated linguistic structures and nuance.
* **Integrate domain-specific knowledge:** By integrating domain-specific knowledge, sentiment analysis' accuracy can be increased. We advise using domain-specific knowledge, such as movie genre, director, and actors, in future sentiment analysis projects on movie reviews to improve the analysis's accuracy. You can accomplish this by:
* Using embeddings or models that have been specially created for movie reviews.
* Incorporating metadata into sentiment analysis models, such as movie genre, director, and actors.
* **Analyze the context of the sentiment:** Sentiment analysis should be considered in the context of the review to properly interpret the sentiment. We advise future sentiment analysis initiatives on movie reviews to assess sentiment in context by accounting for the following elements:
* The review's tenor and vocabulary.
* The genre and style of the movie being reviewed.
* The target audience and their preferences.

Overall, the accuracy and use of sentiment analysis in the context of movie reviews will be enhanced by the suggestions made here. Sentiment analysis can be a useful method for analyzing movie reviews and gaining insights into audience preferences by utilizing high-quality training data, sophisticated machine learning algorithms, domain-specific knowledge, and evaluating sentiment in context.

1. **TOOLS/PLATFORM USED**

**MINIMUM HARDWARE REQUIRED:**

**Web Application:**

* RAM: 4 GB RAM or above
* Processor: ANY

**SOFTWARE CONFIGURATIONS:**

**CLIENT: -**

* Platform used: PyCharm, Jupyter.
* Front - End design: Streamlit

**SERVER: -**

* Coding: Python, Machine Learning, HTML.

**HARDWARE BEING USED:**

* RAM: 4 GB RAM
* Processor: Intel Pentium
* Processor Speed:1.0GHZ
* Hard Disk:1 TB hard disk

**10. Required Libraries**

|  |
| --- |
| cleantext==1.1.4 |
| |  |  | | --- | --- | |  | click==8.1.3 | |  | colorama==0.4.6 | |  | cycler==0.11.0 | |  | et-xmlfile==1.1.0 | |  | Flask==2.2.3 | |  | importlib-metadata==6.0.0 | |  | itsdangerous==2.1.2 | |  | Jinja2==3.1.2 | |  | joblib==1.2.0 | |  | kiwisolver==1.3.2 | |  | MarkupSafe==2.1.2 | |  | matplotlib==3.4.3 | |  | nltk==3.8.1 | |  | numpy==1.21.4 | |  | openpyxl==3.1.2 | |  | Pillow==8.4.0 | |  | pyparsing==3.0.5 | |  | python-dateutil==2.8.2 | |  | regex==2022.10.31 | |  | six==1.16.0 | |  | textblob==0.17.1 | |  | tqdm==4.65.0 | |  | Werkzeug==2.2.3 | |  | zipp==3.15.0 | |

**11. DATA MODELS**

**11.1 Site Map:**

Diagram

Description automatically generated

**Fig 10.1: Site Map**

**11.2 User Activity Diagram**

Diagram

Description automatically generated

**Fig 10.2: User Activity Diagram**

**12. PROJECT SCHEDULING**

**12.1 PERT CHART:**

Task 6.1

Task 6.1

Task 4.2

Task 4.1

Task 7

Task 6

Task 6

Task 5

Task 4

Task 3

Task 1

Task 2

**Critical Path**

**Normal Path**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Task Name** | **Start Date** | **Duration** | **End Date** |
| **1** | **Requirement specification** | **23-01-23** | **5 days** | **27-01-23** |
| **2** | **Research and Analysis** | **28-02-20** | **8 days** | **4-02-23** |
| **3** | **Performed data pre -processing and EDA** | **5-02-20** | **20 days** | **24-02-23** |
| **4** | **UI - Design** | **25-02-23** | **10 days** | **6-03-30** |
| **5** | **Coding: Train various models for sentiment analysis** | **7-03-20** | **30 days** | **6-04-20** |
| **6** | **Testing** | **7-04-23** | **5 days** | **12-04-23** |
| **7** | **Implementation** |  |  |  |

**Fig 11.1: Pert chart**

**12.2 Gantt chart:**

We estimated the number of days for each task as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Task Name** | **Start Date** | **Duration** | **End Date** |
| **1** | **Research and Analysis** | **23-01-23** | **5 days** | **27-01-23** |
| **2** | **Requirement specification** | **28-02-20** | **8 days** | **4-02-23** |
| **3** | **Performed data pre -processing and EDA** | **5-02-20** | **20 days** | **24-02-23** |
| **4** | **UI - Design** | **25-02-23** | **10 days** | **6-03-30** |
| **5** | **Coding: Train various models for sentiment analysis** | **7-03-20** | **30 days** | **6-04-20** |
| **6** | **Testing** | **7-04-23** | **5 days** | **12-04-23** |
| **7** | **Implementation** |  |  |  |

**Fig 11.2: Gantt chart**

**13. Software Requirements Specification (SRS)**

**Introduction**

The following subsections of the SRS document provide an overview of the entire SRS.

**A. Purpose**

The purpose of our project is to provide sentiment for the movie review whether it is positive, negative, and neutral. Apart from that this application also provide basic information about the movie such as movie poster, Rating, Genre and so on. This application also have a data insight section where user can find all the insight about the dataset.

**B. Scope**

The application will be used by General public, Movie production company, Movie Review Website, social media Company.

**C. Benefits**

This application reduces the complexity, reduces a time, make the process comfortable and most important part it provides a sentiment for the whole dataset in minimum time.

**Functional and** **Non-Functional Requirement:**

* To assess the reviews and ascertain the general sentiment, the application deployed a reliable and accurate sentiment analysis method.
* Based on the findings of the sentiment analysis, the application can categorize each review as either positive or negative.
* Based on the ratio of favorable and negative evaluations, the application assigns each movie a final score.
* Users can enter a movie title, check the results of the sentiment analysis, and gain data insight through the application's user-friendly interface.
* Even with a lot of data, the application performs well and produces reliable sentiment analysis results.

**Non-Functional Requirement:**

* Accuracy: The sentiment analysis algorithm should be very accurate at figuring out how each review feels.
* Performance: Even when dealing with a large amount of data, the application should run smoothly and produce results quickly.
* Scalability: The application should be able to accommodate an expanding user and review base without experiencing performance issues.
* Reliability: There should be little to no downtime or errors, and the application should always be accessible to users.
* To serve a diverse user base, the application should be compatible with a range of platforms, gadgets, and operating systems.

**14. Findings**

**14.1 Data Insight**

**Chart, bar chart

Description automatically generated**

**Fig 14.1:** This bar graph shows the distribution of positive and negative reviews, and the dataset is highly balanced as it has an equal number of positive and negative reviews.

**Chart, bar chart

Description automatically generated**

**Fig 14.2:** This graph displays the most common Positive words that are found in the dataset.

**Chart, bar chart

Description automatically generated**

**Fig 14.3:** This graph displays the most common Negative words that are found in the dataset.

**Chart, bar chart, histogram

Description automatically generated**

**Fig 14.4:** This histogram illustrates the number of words in the reviews.

**Text

Description automatically generated**

**Fig 14.5:**  Word-Cloud for most frequent positive words.

**Text

Description automatically generated**

**Fig 14.6:**  Word-Cloud for most frequent negative words.

**15. SYSTEM SECURITY MEASURES**

**Secure and Robust Programming:**

Many system failures and security vulnerabilities arise at the programming level. These can often be attributed to inadequate handling of exceptional situations, poor understanding of the details of the programming language in use, incomplete descriptions of the interfaces between components, and insufficient care in the treatment of concurrency and threading issues.

In this system, special care has been taken while performing the programming for the software so that the software doesn’t get corrupted or meet any fatal error while running.

Use secure connections: Whenever you are making external API requests or downloading files, it is important to use secure connections. You should always use HTTPS instead of HTTP to ensure that the data being transferred is encrypted.

Keep dependencies updated: Keep all dependencies used by the code updated to their latest version to ensure that any security vulnerabilities are patched.

Avoid hardcoding secrets: Avoid hardcoding sensitive information such as API keys, credentials, or passwords in the code. Use environment variables or configuration files to store such information.

**16. TESTING**

**16.1 Testing Techniques and Testing Reports**

**Unit Testing:**

Unit testing, also known as component testing, refers to tests that verify the functionality of a specific section of code, usually at the function level. In an object-oriented environment, this is usually at the class level, and the minimal unit tests include the constructors and destructors.

**Testing Objectives:**

* Testing is the process of executing a program with the intention of finding an error.
* A good test case is one that has a high probability of finding a yet undiscovered error.
* A successful test is one that uncovers a yet undiscovered error.

**16.2 TEST REPORTS FOR UNIT TEST CASE:**

**Test case 1:**

If user enters the wrong Movie name or misplead the name and left the input field blank so system will handle this error by display movie not found.

Background pattern

Description automatically generated

**Fig 16.1: Test Case 1**

**Test case 2:**

In the second test case we checked the accuracy of the model by analyze the dataset that was pre labeled as positive or negative sentiment and model showed a great accuracy. We compare the sentiment and analysis column result, and it analyzed the correct sentiment of the reviews.

Graphical user interface, text

Description automatically generated

**Fig 16.2: Test Case 2**

**17. REPORT**

Graphical user interface, text

Description automatically generated with medium confidence

**Fig 17.1:** This is the front Page of the application. Where you will find two expanders namely Analyze reviews and Data. In the other section you will find data insight section.

Background pattern

Description automatically generated

**Fig 17.2:** These are the input field for movie tittle, reviews and cleaning the data.

A picture containing text, screenshot, screen

Description automatically generated

**Fig 17.3:** When user enter the movie tittle application will extract all the information about the movie such as Genre, Rating, Actors name and so on therefore with the help of all the details user will find out whether the movie is worth watching or not.

Graphical user interface, text

Description automatically generated

**Fig 17.4:** This is the important section of the application where user can enter their reviews and system will tell whether it is positive or negative review by generating the polarity score. If the polarity score is in negative that mean given review is negative else it will be positive.

Graphical user interface, application

Description automatically generated

**Fig 17.5:** In this section user clean their text like removing html tags, special characters, stop words.

A screenshot of a computer

Description automatically generated with medium confidence

**Fig 17.6:** Here user can browse the dataset that they want to analyze, and it can handle dataset up-to 200MB. The best part of the section is user can download the processed dataset.

**18. References:**

* Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Proceedings of the 12th International Conference on World Wide Web (WWW ’03), 519–528. <https://doi.org/10.1145/775152.775226>
* Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the 8th International AAAI Conference on Weblogs and social media (. Retrieved from <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109>
* Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-145. https://doi.org/10.1561/1500000011
* Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 142-150. https://www.aclweb.org/anthology/P11-1015/.
* Natural Language Toolkit. (n.d.). TextBlob: Simplified Text Processing. Retrieved March 27, 2023, from <https://textblob.readthedocs.io/en/dev/>
* Kaur, H., & Mittal, N. (2020). Sentiment analysis of IMDB dataset using machine learning and deep learning approaches. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). <https://doi.org/10.1109/icccnt49239.2020.9225192>
* Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. <https://doi.org/10.2200/s00416ed1v01y201204hlt016>
* Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering, 23(10), 1498–1512. <https://doi.org/10.1109/tkde.2010.259>
* Liu, B. (2012). Sentiment analysis and subjectivity. Retrieved from <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>.
* Roul. (2021). Sentiment analysis - lexicon models vs machine learning. Nerd For Tech. Retrieved from https://medium.com/nerd-for-tech/sentiment-analysis-lexicon-models-vs-machine-learning-b6e3af8fe746.

**19. Appendices:**

**Appendix A: Dataset Description**

In this project, 50,000 movie reviews from the IMDB movie review database were used as the dataset for sentiment analysis. There are an equal number of positive and negative movie reviews in the dataset. Each review is assigned a sentiment score, which ranges from 0 (negative) to 1 (positive) (positive). With 5,000 reviews in each class, the dataset is evenly split between the two classes and is balanced.

**Appendix B: Preprocessing Steps**

Before using the movie review dataset for sentiment analysis, it underwent several preprocessing steps. The actions comprised:

* Lowercasing every text element.
* The deletion of punctuation and special characters.
* Tokenizing each review to represent a single word.
* Removing non-alphabetic tokens and stop words.
* WordNetLemmatizer is used to lemmatize each token to its original form.

**Appendix C: Feature Extraction**In this project, sentiment analysis was conducted using two feature extraction techniques:  
A matrix of the frequency of each word in the dataset is known as a "bag of words."  
TF-IDF: A matrix that illustrates the relative weights of each word in the dataset in relation to each other.

**Appendix D: Model Selection**On the movie review dataset, several models for sentiment analysis were tested. The models tried out consist of:

* VADAR.
* TextBlob.
* Logistic Regression.
* AFINN.

**Appendix E: Evaluation Metrics**Accuracy, precision, recall, and F1 score were the evaluation metrics used to evaluate the performance of the sentiment analysis model. A 10-fold cross-validation was used to calculate the metrics.

**Appendix F: Example Predictions**Following are a few examples of movie reviews along with the model's predicted sentiment scores:

Graphical user interface, text

Description automatically generated

**Appendix G: Source Code**

**Link to the Source code:**

[jaswindersingh1998/Movie\_Analyzer: This app provide details about the movie and also provide the sentiment on the reviews. (github.com)](https://github.com/jaswindersingh1998/Movie_Analyzer)

**Link to the Application:**

[Streamlit](https://movie-senti1998.streamlit.app/)