**Sentiment Analysis on Movie Reviews**

Manideep Charugundla (11479531) Hari Prasad Donthi (11471392)

GitHub link: <https://github.com/Manideepcharugundla/NLP-PROJECT>

[**https://unt.zoom.us/rec/share/Z86FpjenBOEHg6Pi\_FUelorl2feS0HM2BuSbp8OLpp2TUddNMaik6Jj9FARR3vi-.47l7Q\_1S99DGn9Wl**](https://unt.zoom.us/rec/share/Z86FpjenBOEHg6Pi_FUelorl2feS0HM2BuSbp8OLpp2TUddNMaik6Jj9FARR3vi-.47l7Q_1S99DGn9Wl)

**Introduction:**

In this present people share their views on sites. Depending upon feedback of the users, we get the overview or quality of product. People depend upon these feedback or reviews to buy the product or watch the movie. The amount of data is very huge it is difficult to analysis for human manually. We can analysis the data using sentimental analysis. It is divided into two types; one is knowledge based and machine learning approach.

**Background**

**Motivation:**

People generally see the reviews before watching any movie, it plays major for person to watch movie. There are different reviews on different sites, it is difficult for a person to decide whether to watch or not because there are positive and negative reviews on it. For a human it is impossible to make decision. To overcome this problem, we have can sentimental analysis on the movie reviews, it makes the decision easily. People share their thoughts or opinions on social media through comments, posts, tags. We can use the machine learning algorithms to analyze the data. There are different algorithms like navies bayes, SVM. For example, consider social network sites like Instagram, YouTube, fb these sites have huge users base by this there will be huge traffic on the sites and these sites must manage the data of different users coming across the world.

In knowledge-based analysis it mainly depends upon the emotions which require huge dataset which shows the different opinions. The other technique is machine learning approach which make use of training set, it is easier than knowledge based. In this we will having the positive or negative words.

Sentimental analysis is used compare the customer and user depending upon the point of view on movie with the feedback of users. One can easily identify the product is good or not by reviewing the comments or feedback. The major steps in sentimental analysis are preprocessing, extraction, selection, and model.

**Significance:**

Sentimental analysis is commonly as opinion mining which handles with opinions and emotion from text. It is done by using NLP i.e., natural language processing, text, and computational analysis. There is huge development of web technology by this we will getting the huge data from the sites.

**Objectives:**

1. Preliminary sentiment on Reviews
2. Nltk
3. Machine Learning Algorithm
4. Visualization

**Your Model**

**Flowchart:**

Chart, diagram

Description automatically generated

Here we started by collecting the data and the data is processed for the sentimental analysis where varied sites provide different evaluations, making it difficult for a person to determine whether or not to watch because there are both favorable and bad ratings. Requirements are gathered and processed and baseline and developed the code for the analysis using the Machine Learning algorithms and Trained and Tested both the train data set and the test data set and get the train accuracy and the test accuracy scores.

**Architecture:**

Diagram

Description automatically generated

Initially we started with data collection and then preprocessing of data. Here cleaning (Includes Removes the single quotes, stop word removal, tokenization and stemming) understanding the data by using data visualization tools like bar graph and pie chart and join the negative, positive, and neutral reviews and show it as images. Then later we used machine learning algorithms to know the accuracy of the models and after sentiment classification and finally results.

**Analysis of data**

**Features:**

There are different approaches which helps in words into numerical.

Bag of words: Consider an example, there is dataset consists of all unique words in it. We count all unique words which are repeat in that dataset. Which shows the data in vector representation of text.

Word2vec: Word2vec (Word to Vector) is a two-layer neural net that measures text. Its feedback is a text corpus, and its yield is a bunch of vectors: include vectors for words in that corpus. The 50-D space can be pictured by using classical strategies to decrease the vectors to two-dimensional information that can be plotted.

**Dataset:**

we start by depicting the dataset, what the singular sections mean, and how the information was created. Then, we plunge a level further, present the information credits, and mention some underlying observable facts about the information that we have. From here, we conceptualize conceivable use cases that we could settle with the current dataset and survey the information necessities and true ramifications of every potential use case.

The dataset is contained tab-isolated records with phrases from the Rotten Tomatoes dataset. The train/test split has been safeguarded for the reasons for benchmarking, however the sentences have been rearranged from their unique request. Each Sentence has been parsed into many expressions by the Stanford parser. Each expression has a Phrase-Id. Each sentence has a Sentence-Id. Expressions that are rehashed (like short/normal words) are just included once in the information.

**Implementation**

Sentiment analysis is additionally prominently known as assessment analysis or assessment mining. The key thought is to utilize methods from text investigation, NLP, Machine Learning, and semantics to remove significant data or information focuses from unstructured text. This thusly can assist us with inferring subjective yields like the general sentiment being on a positive, unbiased, or negative scale and quantitative yields like the sentiment extremity, subjectivity, and objectivity extents.

Predictive Task: The principle point of this undertaking is to recognize the basic opinion of a film survey based on its text-based data. In this task, we attempt to order if an individual loved the film dependent on the audit they give for the film. This is especially valuable in situations when the maker of a film needs to quantify its general exhibition utilizing surveys that pundits and watchers are accommodating the film. The result of this undertaking can likewise be utilized to make a recommender by giving suggestion of motion pictures to watchers based on their past surveys. One more use of this undertaking is finding a gathering of watchers with comparative film tastes (likes or aversions). As a piece of this venture, we plan to concentrate on a few element extraction methods utilized in text mining for example catchphrase spotting, lexical proclivity, and factual techniques, and comprehend their pertinence to our concern. Notwithstanding highlight extraction, we additionally investigate diverse grouping procedures and investigate how well they perform for various types of element portrayals. We at last reach an inference with respect to which mix of element portrayals and characterization strategies are generally exact for the current prescient undertaking.

The decision tree is worked by recursively apportioning the element space into two sections. At each progression, the split that works on the blunder of the tree on the preparation information is utilized, and this insatiable procedure proceeds until a tree of wanted size is created. For new information, a mark is anticipated by following the parts of the tree from the root hub as per the upsides of the elements of the new information. Choice trees enjoy the benefit of being effectively interpretable, as the twofold tree design can be addressed in a drawing, and a person can follow the branches down the tree as indicated by the information factors.

General Sentimental Analysis

An essential framework that performs with a degree of sentiment analysis on movie reviews dependent on motion pictures with following advances:

a. Fetch data from movie reviews.

b. With Feature selection and Word Features, get both positive and negative reviews. Movie reviews and utilizing Classifier we can examine reviews which are positive and negative

The proposed a framework to foresee Box-office assortment dependent on Sentiments of film survey. They have utilized Viewer suppositions are utilized as information factors notwithstanding indicators and three AI based calculations (fake neural organization, relapse tree, and backing vector relapse) were utilized to get non-straight connection between the movies and its assortment indicators.

**Limitations:**

the fundamental issues that exist in the current strategies are powerlessness to perform well in various spaces, lacking exactness and execution in sentiment analysis dependent on inadequate named information, ineptitude to manage complex sentences that require more than sentiment words and basic dissecting.

This stage utilizes a ML-based characterization calculation for classification of users’ reviews. The work of survey arrangement is to sort the user's reviews as negative and positive. In this work, we have utilized the Naive Bayes (NB) classifier as it is both vigorous and exact classifier and gave better outcomes on adaptable datasets in contrast with the other benchmark characterization calculations. Besides, the NB classifier is extremely simple to utilize and thusly has various applications in text order issues.

Initially, the NB classifier takes as information the element vector portrayal of film surveys alongside the names, to order the audits. Likelihood/shot at a term given specific classification (negative or positive) in the survey report is registered dependent on events a term occurred with that class. Here, the term is unigram, bigram, or trigram. To classify another case of survey, the possibility/likelihood of each unigram (single word), bigrams (2 words), and trigrams in the audit given objective name (+ve) is figured, and afterward generally speaking likelihood/shot at the survey given objective mark (+ve) is dictated by taking result of all terms probabilities and the likelihood of the objective class/name (+ve). Similarly, the shot at the survey given objective mark (−ve) is assessed.

**Results:**

We have divided the phrase data and added sentiment according to it.

Graphical user interface, text, application

Description automatically generated

If we plot the graph, we get the result as

Chart, bar chart

Description automatically generated

As we can see that sentiment 2 which has highest frequency which is about 80000.

Chart, pie chart

Description automatically generated

Preprocessing the Data set like remove single quotes, word tokenization using text-to-word-sequence, stop word removal, Stemming.

After that we gathered the reviews based on negative, somewhat Negative, neutral, positive, somewhat Positive.

**Negative**

Text

Description automatically generated

**Somewhat Negative**

**Text

Description automatically generated**

**Neutral**

**Text

Description automatically generated**

**Positive**

**Text

Description automatically generated**

**Somewhat Positive**

**Text

Description automatically generated**

We will get all movie data by neutral, positive, and negative. By performing Naive bayes model we get accuracy of

Table

Description automatically generated

Similarly, if we perform the SVM model on reviews we get the accuracy as

Table

Description automatically generated

For train data on decision tree, we get the accuracy as 96%, where tree test accuracy score is 57%.

Table

Description automatically generated

**Project Management:**

Meeting Deadlines

Description: Due dates for the assigned work.

Any work that has been assigned to an individual or a group, is unable

to get completed in the given deadline results in the drop of grades and is considered as a

high-risk factor.

Plan to minimize risk: Proper guidance and communication among the team and

completing the work with added prior buffer time to the deadline will meet the

requirement.

**Work Completed:**

**Description:**

This task is done by the two members. We have divided the into equal halves. Data collection, cleaning of data, extraction all the essential details and helping each other for implementation of algorithm of getting the accuracy of data. We have implemented the analysis using different models like naive bayes, svm and decision tree.

**Responsibility:**

Manideep Charugundla

▪ Role: Project Team Organizer, Designer & Project Team Member

▪ Lead Team Member, code implementation, understanding the dataset

▪ Coordinated Meetings, Collaborated in Research, Report Writing and Editing Process voice over

Hari Prasad Donthi

▪ Project Team Member, code implementation, cleaning the dataset

▪ Collaborated in Research, Results, Report Writing and Editing Process

▪ Voice over for the project

**Contributions:**

Person Coding Report Discussions/Leadership

Manideep Charugundla 50 55 50

Hari prasad Donthi 50 45 50

Total 100% 100% 100%

**Conclusion:**

For Naive bayes model we get accuracy of

Naive bayes Train Score is 63%

Naïve bayes Test Score is 58%

For SVM model we get the accuracy as

SVM Train Accuracy Score is 59%

SVM Test Accuracy Score is 57%

For Decision Tree we get the accuracy as

Decision Tree Train Accuracy Score is 96%

Decision Tree Test Accuracy score is 57%.

From the above three models it is clearly shown that the Decision Tree has the best Accuracy for the training set and the test set while comparing with the other models.

**References**

1. [Neethu, M. S., & Rajasree, R. (2013, July). Sentiment analysis in twitter using machine learning techniques. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-5). IEEE](https://d.docs.live.net/0394be62cd2b579d/Desktop/Project%20incremental-2/Sentimental-analysis.docx)
2. [Sahu, T. P., & Ahuja, S. (2016, January). Sentiment analysis of movie reviews: A study on feature selection & classification algorithms. In Microelectronics, Comp](https://d.docs.live.net/0394be62cd2b579d/Desktop/Project%20incremental-2/Sentimental-analysis.docx)
3. [Kuat Yessenov, Sasa, ’Sentiment Analysis of Movie Review Comments’, 2009.](https://d.docs.live.net/0394be62cd2b579d/Desktop/Project%20incremental-2/Sentimental-analysis.docx)