SEMANTIC ANALYSIS ON MOVIE REVIEWS

Group Members:

Jaynam Sanghavi - 270526536

Dikshita Patel - 862605004

Vrushali Shah - 355394220

Introduction -

Movie reviews are crucial in evaluating a film's performance. While numerical or star ratings offer a quantitative assessment of a movie's success or failure, textual reviews provide a deeper qualitative understanding of different aspects of the film. Sentiment analysis, a significant subject in machine learning, is closely linked to natural language processing and text mining. It aims to extract subjective information from textual reviews, such as the reviewer's attitude towards various topics or the overall polarity of the review. By using sentiment analysis, we can determine the reviewer's state of mind while providing the review and comprehend their emotions, whether they were happy, sad, or angry.

Semantic analysis has become more accurate and efficient, making it a valuable tool for movie producers, directors, and marketers to gain insights into their target audience's preferences and opinions.In this project, we seek to use sentiment analysis on a set of movie reviews to identify the overall reaction of the reviewers towards the movie, whether they liked or hated it. We plan to utilize the relationship between words in the review to predict the review's overall polarity.

Problem Context:

Sentiment analysis is a process of identifying and extracting subjective information from text, such as opinions, attitudes, and emotions. One common application of sentiment analysis is analyzing movie reviews to determine whether the sentiment expressed is positive, negative, or neutral.

Movie reviews can provide valuable insights into the quality and popularity of a film, and sentiment analysis can help automate the process of analyzing these reviews. Sentiment analysis can be used to identify key features of a movie that viewers liked or disliked, as well as overall trends in sentiment across multiple reviews.

To perform sentiment analysis on movie reviews, machine learning algorithms are typically used to classify text as positive, negative, or neutral. These algorithms rely on training data, which consists of labeled examples of text that have already been classified by humans as positive, negative, or neutral.

Overall, sentiment analysis can be a useful tool for filmmakers, movie critics, and audiences to gauge the reception of a film and understand the key factors driving positive or negative sentiment.

Problem Statement:

Sentiment analysis on movie reviews is to provide insights into the reception of a film and the factors driving positive or negative sentiment, helping filmmakers and studios make more informed decisions about future productions and marketing strategies.
Our aim is to develop a model that can accurately classify the sentiment of each review as positive, negative, or neutral. The model will be able to identify the key features of the text that contribute to the sentiment expressed and be able to generalize to new and unseen reviews.

This analysis would help to understand the audience preferences. It would also help directors or producers to identify areas of improvments.

Objective / Outcome -

The objective of performing sentiment analysis on movie reviews is to determine the overall reaction of the reviewers towards the movie, whether they liked or disliked it. The outcome of sentiment analysis is a quantitative assessment of the subjective information extracted from the textual reviews, such as the reviewer's attitude towards various aspects of the movie or the overall polarity of the review. By analyzing the sentiment of the reviews, we can gain insights into the audience's opinions and emotions, which can be useful for movie producers and marketers to improve their movies' quality and marketing strategies. Sentiment analysis can also help moviegoers to make informed decisions about which movies to watch based on the general consensus of the audience's reactions.

Research Scope -

Our research scope includes the following key areas::

Data Collection Scope: The research will use a publicly available dataset of movie comments collected from various online platforms such as IMDB, Rotten Tomatoes, and Metacritic. The dataset will be preprocessed and cleaned to remove any irrelevant or misleading data.

Data Cleaning and Analysis: The study will be divided into two sections. The first stage will consist of a qualitative examination of a sample of movie reviews. The analysis will include classifying the remarks, identifying the stated emotions, and assessing the tone of the words. The second portion of the presentation will be devoted to a quantitative analysis of a bigger dataset of movie comments. To identify the sentiment, subjects, and entities in the comments, NLP methodologies and techniques such as sentiment analysis, topic modeling, and entity identification will be employed.

Sentiment Analysis Models: A IMDB (Internet Movie Database) dataset that contains will be used to train the model. The model will then be used for sentiment classification provided a movie review.

Machine Learning Algorithms: A review of different machine learning Algorithms that can be used to develop a sentiment analysis model. This review will cover algorithms such as naive bayes classifier, K-Means algorithm, bag-of-words model.

Solution Scope:

The solution scope of this project includes the development of a sentiment analysis model that can accurately predict sentiments for movies. The following components will be included in the project:

Data Collection & Preparation:
The reviews that have been gathered from reliable sources are initially stored to create the review database that will be used to construct the model. The next step is to prepare the data for feature selection. Every review in the database is pre-processed. This entails deleting tab spaces, newlines, splitting the sentences, numbers or digits, punctuation, and stopwords (stopwords are frequently used words like "a" and "an" that the machine learning system can disregard). All the characters are also changed to lowercase. The result is saved in separate text files after this phase is finished. A Python application is used throughout the entire process.

Bag-of-Words Model:
The vocabulary that is being utilized to develop the model is simply depicted by the "Bag-of-Words" model. This was constructed using the review database. The classifier additionally makes use of this vocabulary to pinpoint the characteristics of input before categorizing it into the appropriate group. Natural language processing (NLP) and information retrieval are the foundations of the bag of words (IR). It can be compared to a dictionary that keeps track of different words and maps them to their counts, where counts represent the frequency of occurrence of a term in the dataset used to create the model. It is created using an internal collection called a counter ().The pre-processed words (tokens) used for this are stored along with their frequency information. The outcome is saved in a straightforward text file for future use.

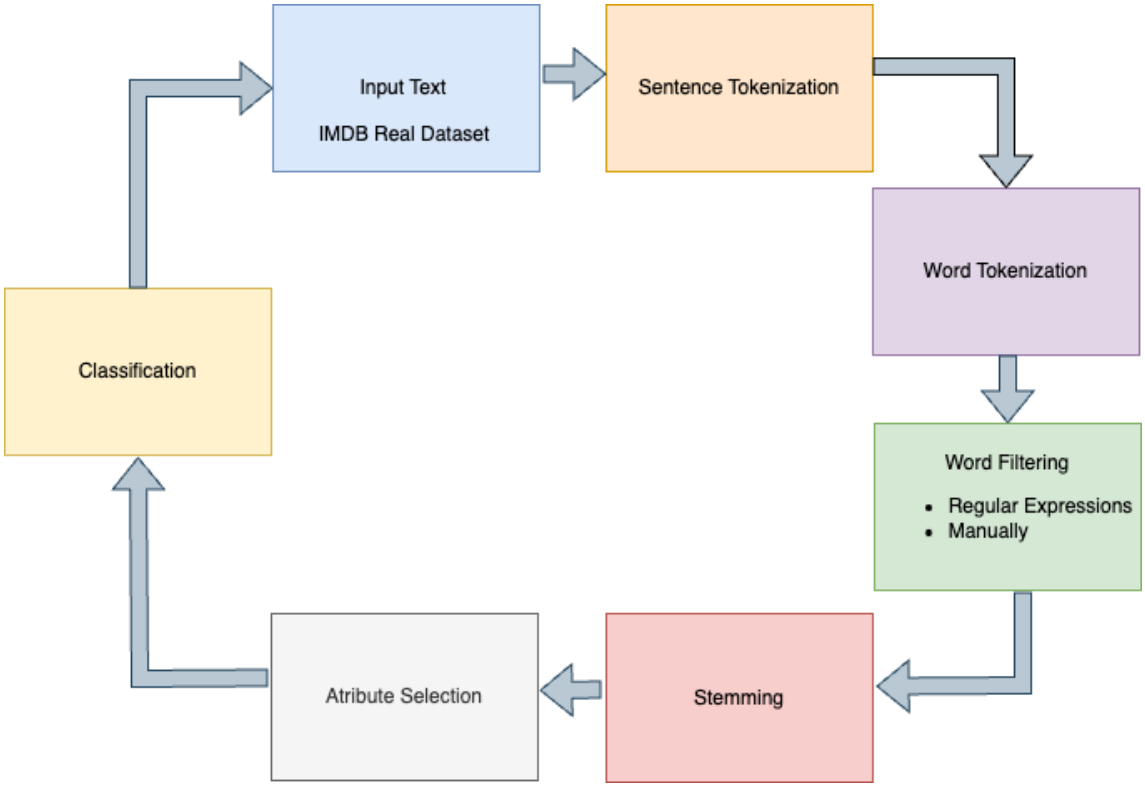
Classification:
The classifier model is trained and tested using both the vocabulary and review databases simultaneously. The datasets are divided into training and testing portions in a 60:40 ratio. The classifier performs as a binary classifier and will categorize the review or remark as positive or negative and assign a rating to each review based on the words that the BoW model matched. The final rating for a given movie is then computed using the average of all the individual comments made about it.

Naive Bayes Algorithm:
A straightforward but powerful classification algorithm is naive bayes. The Naive Bayes method is a popular approach for classifying documents. By applying the Department of Computer Engineering Semester VIII 2015-19 Batch Page 12 combined probabilities of words and categories, the main idea is to estimate the probabilities of categories given a test document. The assumption of word independence is what makes this model naïve. This assumption's simplicity makes the Naive Bayes classifier's calculation much more effective. To represent the real classes as positive (1) and negative (0) in binary representation for implementation purposes, the linear arrays "ytrain" and "ytest" are created. 'Scikit"s built-in functions are used to train and test the Naive Bayes classifier.The accuracy is assessed by comparing the actual and projected outcomes and using the built-in ROC curve measure.

Support Vector Machine (SVM):
The best text classification technique relies on a discriminative classifier called support vector machines (SVM). A statistical technique for classifying data is the support vector machine. Based on the computational learning theory's structural risk minimization principle, SVM finds a decision surface to divide the training data points into two classes and bases its judgments on the support vectors that have been chosen as the only useful components in the training set. This is used to categorize sentiment.

Collecting Reviews for a Movie:
The film for which a rating is required is chosen. You may look up reviews for this movie on reputable websites like IMDB, Metacritic, Rotten Tomatoes, and TMDB. Using a program named "Scrape Storm," the reviews are scraped from various websites. They are taken out and stored in a local database as CSV files, which are then fed into the classifier to provide a final movie score.

Project Methodology:



The execution steps that are shown in the above figure could be summarized as the following:

There is a step-by-step algorithm to perform sentiment analysis of movie reviews using sentence tokenization, filtering, stemming, and classification:

Here is a step-by-step algorithm to perform sentiment analysis of movie reviews using sentence tokenization, filtering, stemming, and classification:

Sentence Tokenization: The first step is to break the review into individual sentences using a sentence tokenizer. This helps to analyze the sentiment of each sentence independently.

Filtering: The second step is to remove any irrelevant words or noise words that do not contribute to the sentiment analysis, such as stop words (e.g., "a", "an", "the"), punctuation, and special characters. This can be done using a pre-built stop word list or a custom list of words specific to the movie domain.

Stemming: The third step is to reduce the words to their root form or stem, which helps to reduce the number of unique words and make sentiment analysis more efficient. This can be done using a stemming algorithm, such as Porter Stemming or Snowball Stemming.

Feature Extraction: The fourth step is to extract relevant features from the preprocessed text, such as word frequencies or n-grams, which represent the sentiment of the text. This can be done using various techniques, such as Bag-of-Words, TF-IDF (Term Frequency-Inverse Document Frequency), or Word Embeddings.

Classification: The final step is to classify the sentiment of the review into positive, negative, or neutral, based on the features extracted in step 4. This can be done using a machine learning algorithm, such as Naive Bayes, Support Vector Machines, or Random Forests.

Overall, the sentiment analysis algorithm using sentence tokenization, filtering, stemming, and classification aims to analyze the sentiment of a movie review by breaking it into sentences, removing irrelevant words and noise, reducing the words to their root form, extracting relevant features, and classifying the sentiment as positive, negative, or neutral. The accuracy of the algorithm depends on the quality of the preprocessing, feature extraction, and classification.

Resources Required

Hardware Requirements for Development :-

- Computer/Laptop

Software Requirements for Development :-

- Python 3.5+
- Python NLP Kit
- 10/8/7/Vista/2003/XP (64-bit)
- 2 GB RAM minimum.
- Links for extracting datasets from TMDB, IMDB, etc.
- Scrape Storm (Desktop Application)

Hardware Requirements for Deployment :-

- Computer/Laptop

Software Requirements for Deployment :-

- Internet Browser

References -

1. Tirath Prasad Sahu, Sanjeev Ahuja, “Sentiment analysis of movie reviews: A study on feature selection & classification algorithms.” IEEE, 2016
2. Mr. B. Narendra, Mr. K. Uday Sai, etc. “Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies.” I.J. Intelligent Systems and Applications, 2016
3. Jyotika Yadav, “A Survey on Sentiment Classification of Movie Reviews.” IJEDR, 2014