**Title**

**Group 1**

**Team Members:- Daixuan Chen, Vyas Pranali, Wentao Gao Amy, Nimmikrishna Babu**

**Introduction**

Movie reviews are vital for determining the reception of a movie and to decide it as a success or failure. As consumers of the visual media product, they have helped define the various genres which in turn help content makers to produce content that cater the general need. Movies when classified under genre are expected to follow certain aspects and attributes that define the genre and movie viewers when selecting to watch a movie under these specific genres would want it to satisfy these attributes failing to which will lead to failure of the movie. Often expectations from a genre are changed or new genres are formed based on viewer reviews. So analysis of movie reviews by viewers are very much integral part to movie making to understand the public emotions.

Movies reviews can be analyzed to study the public take on the movie after its release to be positive or not. Based on the sentimental analysis and other evaluation metrics visual media platforms such as Netflix have built strong recommendation systems.

In this paper we are trying to find out what aspect of the movie is majorly contributing to the success or failure of a movie under a particular genre. By this content makers when making movies under a particular genre can focus on improving the features or attributes that can gather positive feedback from viewers.

**Literature review**

Paper Title:- Movies Reviews Sentiment Analysis and Classification

Authors:- Mais Yasen, Sara Tedmori

In this paper the authors are building a sentiment analysis model that classifies movie reviews. They employ tokenization to produce word vectors from input strings, extracts root words using stemming, essential words are extracted using feature selection and finally labelled as positive and negative. The database chosen is IMDB for training and testing and results are compared sing different metrics.

The authors are primarily focused on comparing the performance of different algorithms. To evaluate the proposed model, eight different well- known classifiers were run on the same training and testing datasets : NB(Naïve Bayes), BN(Bayes Network), DT(Decision Tree), KNN(K-nearest neighbors), RRL(Ripper Rule learning), SVM(Support Vector Classifier), RF(Random Forest), and SGD(Stochastic gradient descent) classifiers in SA. The authors have concluded that RF has better efficiency over other classifiers.

With the background knowledge of these studies we are further aiming to do a sentimental analysis on the reviews by categorizing the contents of the reviews into categories such as comments on Direction, acting, plot of the movie, sound-effects, VFX etc. We will then analyze these data to classify them into positive and negative remarks and study which of the categories are more often commented and deciding the overall tone of the review of the movie.

**Research Question**

The research question for the project is: to find out what aspect of the movie is majorly contributing to the success or failure of a movie under a particular genre?

**Data Collection**

The data for this project is collected from IMDB movie review site. We used Selenium webdriver to navigate the link and find elements by id, name and CSS selectors and saved the data into a csv file. For this report reviews for movie Tenet was collected, cleaned to remove any non-alphanumeric characters, links etc from the title of the review and the review text content and preprocessed to remove stopwords by using the NLTK database of stopwords. After that the data is tokenized, by which the review text is broken down into the smallest unit in a sentence called tokens.

This was followed by pos-tagging by which the words in the corpus was categorized in correspondence with a particular part of speech, depending on the definition of the word and its context. Finally the reviews are stemmed using the PorterStemmer in the NLTK library.