# **Sentiment Analysis and Word Cloud Generation for IMDb Reviews**

# **Introduction**

Sentiment analysis, also referred sometimes as opinion mining, a process which uses natural language processing (NLP) to extract, identify and categorize different views and opinions provided in text format [1]. It involves the use of data mining, machine learning, artificial intelligence and computational linguistics to achieve the task. Its helps to discern whether the author of the text had good, negative or neutral attitude towards a given topic or area of discussion [2].

Sentiment analysis helps the businesses to provide business objective review and insights, feedback for better product portfolio and customer satisfaction. Generally, there are multiple stages that the analysis goes through before giving the result. The first stage is the preprocessing stage, during which the sentiment analysis will identify key words which highlights the core message of the text. Then comes the keyword analysis which further analyse the extracted keywords and given them a sentiment score. The score is a measurement scale that indicates the emotional element in the sentiment analysis system [3].

The two commonly used approaches for sentiment analysis includes rule-based sentiment analysis and automated or machine learning sentiment analysis. In the first approach sentiment analysis is done based on manually created rules, whereas in the second approach the analysis depends on machine learning algorithms. There are approaches which is a hybrid of these two approaches [4 ].

Some of the challenges include neutral sentiments, unclear or ambiguous text, unclassified language like emojis, fake reviews etc [5].

# **Scope**

The project aims to analyse the IMDB reviews of one the movies. Identify the overall user satisfaction and examine the positive and negative comments of the reviewers in the form of wordcloud. This will help in analysing the user satisfaction in the form of words and the phrases specified on their comments.

# **Objectives**

Through this paper we aimed to review the analysis of a well-known movie from IMDB. We have performed a sentiment analysis on these reviews and generate the word clouds. This helped us in analysing the viewers emotions and visualise most frequently used words and phrases. The reviews were extracted from IMDB. We performed preprocessing steps like text cleaning, tokenizing and stop words removal. This helped us in highlighting key words. The preprocessing included removing numbers, symbols, common stop words and punctuations etc. We also wanted to create word clouds for positive and negative words using positive and negative lists. The outputs that we had envisaged were cleaned data frame, word clouds identifying most frequent words with unigrams and bigrams and observe negative and positive sentiments in visualisation.

1. **Data Collection:**
   1. Extract the reviews from [**The Dark Knight**](https://www.imdb.com/title/tt0468569/?ref_=tt_urv)(2008) using the web scraping method.
2. **Data Preprocessing:**
   1. To make this data suitable for the analysis, clean and preprocess the extracted reviews.
3. **Analyse the Text Data:**
   1. Use TD-IDF (Teram Frequency Inverse Document Frequency) and N-gram analysis to find the most used sentiments and phrases.
4. **Sentiment Analysis:**
   1. Examine reviews for positive and negative sentiments. Visualize them with the wordcloud.
   2. We will also find the polarity score for unigram and bi grams
5. **Visualization:**
   1. Unigram and Bigram: Visualize the frequently appearing words in the reviews with the help of Unigram. Bigram for frequently occurring two words.
   2. Positive and Negative Wordclouds: Create wordcloud for most common and significant positive and negative words present in the reviews.

# **Methodology**

1. **Tools and Libraries:**
   1. The project is implemented using spyder ( conda: base (python 3.11.8)) and Jupyter
   2. Python programming language for implementing the project.
   3. Python Libraries Used:
      1. **Pandas:** for data manipulation
      2. **Requests:** HTTP library, used to request and get an access the web content.
      3. **BeautifulSoup:** for web scrapping
      4. **Re:** For regular expression operations.
      5. **Nltk:** Natural Language Processing library
      6. **Matplotlib:** used for data visualization
      7. **Wordcloud:** for wordcloud generation and visualization

A diagram of a diagram

Description automatically generated

**Fig. 1 Project Architecture**

1. **Data Collection**
   1. We Scrape IMDb reviews for a movie The Darak Knight 2008, by providing the URL: [The Dark Knight (2008) - The Dark Knight (2008) - User Reviews - IMDb](https://www.imdb.com/title/tt0468569/reviews?ref_=tt_urv).
   2. As the name indicates the ‘*requests*’ library, request HTTP server to grant an access to fetch the web page content.
   3. Once the request is approved, the *‘BeautifulSoup’* will iterate over the HTML content and searches for a tags *‘div’* and class *‘text’* or *‘test show-more\_\_control’*. Content under these tags will be extracted.
   4. We will then store the reviews in a list and save them to a text file for further analysis.
2. **Data Preprocessing:**
   1. We will combine all the different reviews and perform below tasks:
      1. In this stage the raw text data cleaned and prepared to ensure that the analysis yields accurate and meaningful insights.
      2. In this process we will eliminate all the unwanted characters, except for A-Z, a-z, 0-9 and ‘spaces. All the ‘, ‘ (commas) will be replaces by ‘space’. Removing the unwanted symbols will reduce the noise in the data and enhances the text clarity.
      3. To ensure that all the words are treated uniformly, we will convert the text to lowercase. For example, "Dark" and "dark" is treat as the same word.
      4. Split the test into words to remove the stopwords and some additional words related to movie (e.g. Batman, Christian Bale, etc.) which will not contribute to analysis.
   2. These steps are performed so the clean text will yield better tokens.
3. **Text Analysis:**
   1. In our project we have used TF- IDF (Term Frequency Inverse Document Frequency) – ‘*TfidfVectorizer*’ to convert the text data into matrix. This will help to find the word appearing in each review. Higher the frequency value, more unique is the word in the document. For example, “absolutely”, “batman”, “outstanding”, etc are unique word in the document 1 and hence has high frequency values of 0.38.

**A screenshot of a calendar

Description automatically generated**

**Fig. 2 Example TF-IDF Matrix**

* 1. We will generate the unigram and visualize the wordcloud to see these frequencies.
  2. Generate unigrams and bigrams and visualize them using word clouds to understand the frequency and importance of different words and phrases.
  3. Positive and negative words are segregated from the words list and visualize them with the wordcloud.
  4. In order to identify the the bigrams we well lemmatize the words with the help of ‘*WordNetLemmatizer*’ from ‘*nltk*’ and use ‘*bigrams()’* function to from the bigrams, which intern can be visualized with wordcloud.

1. **Sentiment Analysis:**
   1. In this method we use ‘*SentimentIntensityAnalyzer*’ from *‘nltk’* to calculate the sentiment scores from single words as well as bigrams.
   2. The sentiment will be saved in the form of dataframe, which will be used to analyse the distribution of sentiments (positive, negative and neutral).
2. **Visualization:**
   1. In this project we use ‘WordCloud’ module from ‘wordcloud’ library to visualize the overall words, most commonly and significant words from the reviews and also, positive and negative words.

# **Result**

After adding some of the most common words like name of the file makers, directors, movies name and other common words, we see that “good” is most frequently used word in the reviews. This word is also associated with positive sentiments of reviewers towards movie and actors’ performance.

A close up of words

Description automatically generated

**Fig. 3 Frequently used words in the reviews**

Positive and negative wordcloud are shown in the images below:

A close up of words

Description automatically generatedA close up of words

Description automatically generated

**Fig. 5 Negative Wordcloud**

**Fig. 4 Positive Wordcloud**

The words “Plot” and “dent” has appeared in the negative wordcloud indicating some of the negative words used while writing the reviews. These are some of the common critiques about the story.

The bigram below shows that most together appearing words in the reviews, which indicates that views have praised major key players of the movies, performance of the actors. There are also critiques about specific elements like the plot and visuals etc.

A close up of words

Description automatically generated

**Fig. 6 Bigram Wordcloud**

Sentiment analysis provided an insight into the emotional tone of the reviews. Whereas sentiment score provides an information on positive and negative sentiments across the different terms like ‘phenomenal’, ‘finest’, ‘sequel’ etc. The sentiment scores for some of the terms, unigrams and bigrams are shown in the screenshot below.

A screenshot of a computer

Description automatically generated

**Fig. 7 Sentiment Score for Unigrams**

A screenshot of a computer

Description automatically generated

**Fig. 8 Sentiment Score for Bigrams**

# **Conclusion:**

The objective of the project is to analyse, extract inferences and sentiments from the IMBD movie reviews with the help of Natural Language Processing Technique. By implementing web scrapping, text processing, sentiment analysis, lemmatization and TF-IDF, we were able to derive the findings.

# **Implications:**

The analysis will provide the feedback to the producers and filmmakers, providing the strengths and weaknesses and areas of improvements required in accordance with their audience. The methodology used in his project is applicable for other customers reviews from different platforms. The approach is scalable for text mining and sentiment analysis for numerous domains.