MI2 Sea Monsters Checkpoint
Group: The Sea Monsters
Group Leader: Audrey Himes Group Members: Alexa Owen, Sharaf Tariq
DS 4002
February 14, 2023

Hypothesis: There is a connection between movie review sentiment and the ultimate rating that the film receives.

Research Question: Can the sentiment of the language used in a movie review be related to a numerical rating from 1 to 10?

Model Approach: We will be using natural language processing and text mining to conduct a sentiment analysis. Our main approaches will be to use the National Resource Council emotion lexicon (nrc) algorithm, afinn analysis, and bing analysis to understand the sentiment within our text data.

Executive Summary: Sentiment analysis will be performed on the chosen IMDb (Internet Movie Database) Movie Review data set. The analysis will relate the sentiment of the review to the numeral rating given by the reviewer. This will be done by putting the numerical ratings into groups (1-3, 4-7, 8-10) and seeing which words are usually associated with each rating group.

Data Set Establishment Details: Our data contains 5,450 reviews for various movies that were released in 2021. All reviews have been sourced from IMDb (Internet Movie Database) and represent the opinions of users of the site. Our data contains five variables to be described in the data dictionary below.

Element/Variable Display Name	**Description**	**Data Type**	**Acceptable Values**
ID	Observation number	integer	Any integer 1-5450
REVIEW	Text review about an IMDB movie	character	Any character string
RATING	Overall numerical rating given by the user	integer	Any integer 1-10
AUTHOR	IMDB review author name	character	Any continuous character string (no spaces)
TITLE	Title heading of the review	character	Any character string

This data set is available on Kaggle [1].

We have begun to explore a few questions related to our data that are guiding our exploratory data analysis. They are listed out below along with the preliminary findings that have come as a result of our exploration.

What scale is used for rating and how will the ratings be divided?

The reviews in this dataset are cast on a scale of 1-10, with 1 being the worst and 10 being the best. We will use these data points to refine our research question so that we can analyze how sentiment of a review can relate to a specific numerical rating scale. Further, the data will be divided based on this scale into three different sets. Based on how the data is balanced, the ratings will be divided into three sets: the first set will contain ratings from 1-3 (Low), the second from 4-7 (Medium), and the third from 8-10 (High).

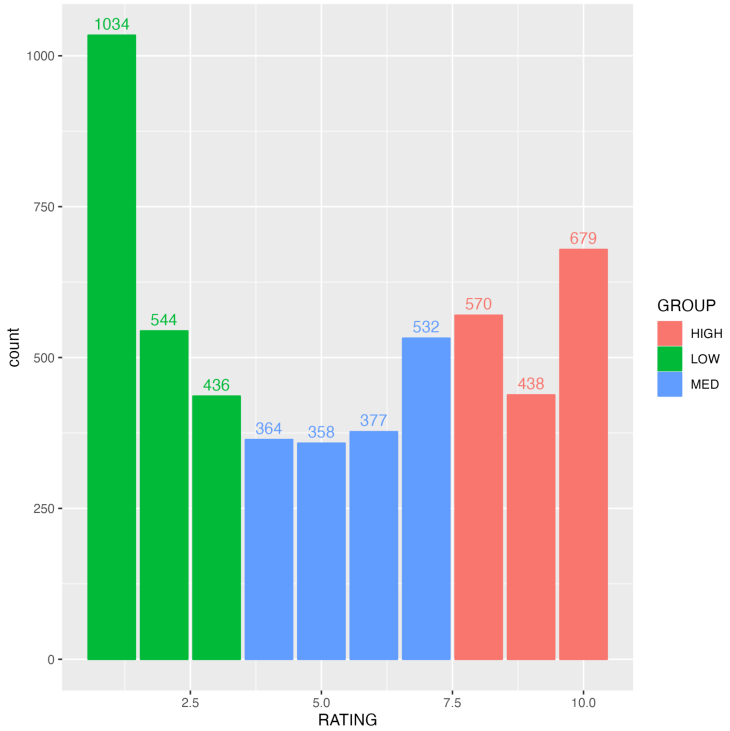
Are these groups balanced?

The groups that we created (Low, Medium, High) contain the following number of reviews:

Low (1-3): n = 2014
Medium (4-7): n = 1631
High (8-10): n = 1687

While there are slightly more reviews in the Low group than in Medium or High, this is not of concern towards understanding and answering our research question.

The below plot is a visualization of the number of words that fall within each numerical rating, colored by the rating groups that we created.



Additionally, the groups contain the following number of individual words:

Low (1-3): n = 158705
Medium (4-7): n = 225348
High (8-10): n = 181749

Similar to the review count, there are slight variations between the total number of words in each rating group. These variations are not of major concern towards answering our research question.

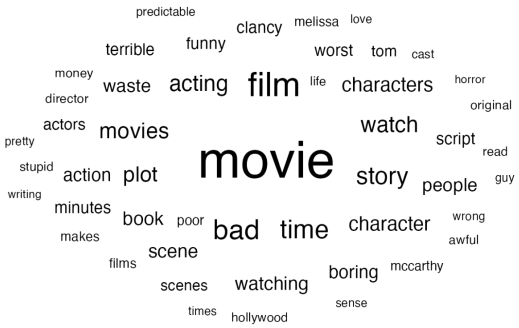
Which variables are relevant for analysis?

We determined that out of the five variables in the data set, ‘REVIEW’ and ‘RATING’ are the most relevant to the research question. We therefore removed all other columns from the data set for the sake of our analysis. Rows containing NA values were removed as well during the data cleaning process. This reduced the number of observations from 5,450 to 5,332.

What words are most prevalent in each set?

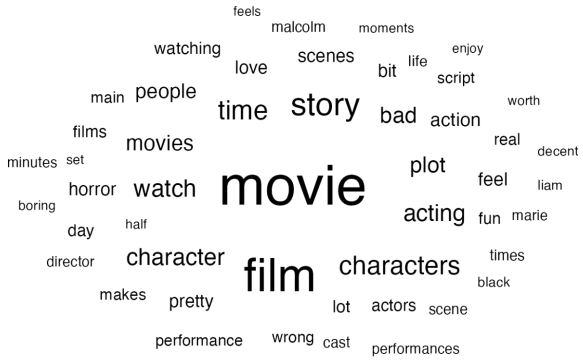
Understanding what words appear most frequently in our data set is vital towards understanding how the sentiment of reviews can affect or generate a numerical rating. To explore this, we created a word cloud for each rating group (Low, Medium, High). In creating the word cloud, we removed all numbers and “stop words” from the data [2]. The word clouds are shown below.

Low:



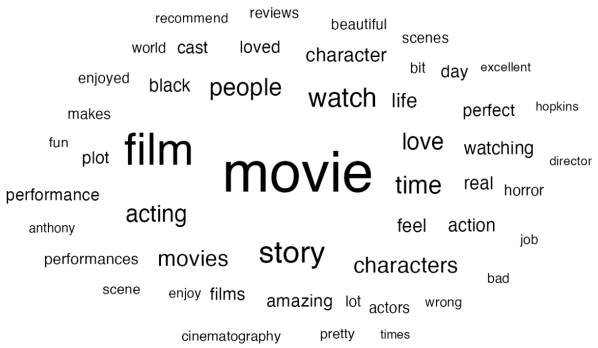
The 5 most common words are “movie”, “film”, “bad”, “time”, and “story”. While we can’t draw any conclusions based on this exploratory analysis, it is in line with our hypothesis that “bad” is one of the most prevalent words in the low rating group.

Medium:



The 5 most common words are “movie”, “film”, “story”, “time”, and “characters”. This does not provide us with any real information surrounding the sentiment of this set, although we can see within the word cloud that there are a mix of positive and negative sounding words.

<u>High:</u>

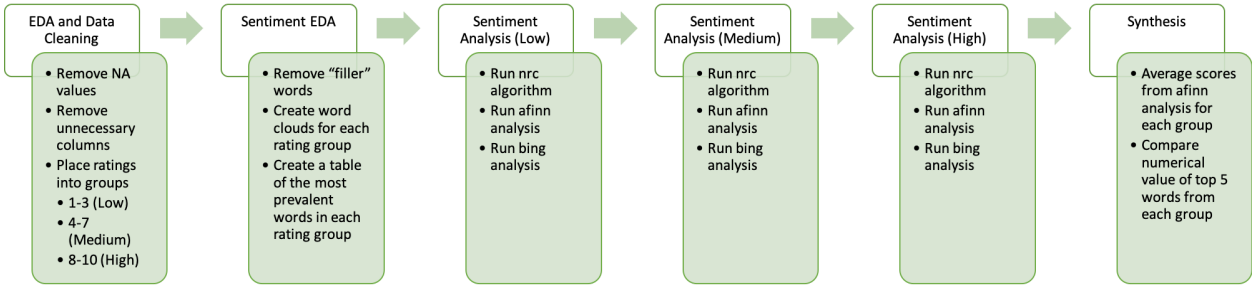


The 5 most common words are “movie”, “film”, “story”, “watch”, and “time”. Similarly to the medium rating set, this does not provide us with a ton of information or context for what the rest of the set is going to look like. However, we can see from the word cloud that there are many words with a positive association such as “enjoy”, “love”, “beautiful”, and “excellent”. This is in line with our hypothesis.

As we can see, many of the most common words in each set are “movie”, “film”, “time”, and “story”. In order to most effectively answer our research question, it may be beneficial to remove these words from our analysis. Through the lens of a sentiment analysis related to movies and film reviews, it is unlikely that these words would change our results. Rather, they could just be seen as unnecessary fluff that stands in the way of efficiently drawing conclusions from our data.

While the above exploratory data analysis is certainly interesting and vital towards forming a rudimentary understanding of the sentiment of our text data, there are still some questions left unanswered that we will explore in future steps in our analysis plan. Specifically, we need to conduct more precise sentiment analysis on each rating group so that we can create some sort of numerical way to understand the sentiment and emotions that lie behind the text data. Additionally, while we can look at a word cloud and have a rough understanding of the types of words that may exist within each rating group, we must conduct further sentiment analysis in order to determine how each word should be characterized without personal bias playing a role.

Analysis Plan:



EDA and Data Cleaning: To isolate the data we will be conducting EDA (Exploratory Data Analysis) and data cleaning in the first step. In this step, NA values will be removed and the number of variables reduced to 'Review' and 'Rating' for our analysis. After isolating the variables, we will group the ratings 1-3, 4-7, and 8-10 so that the ratings and reviews are in individual data sets according to 'Low', 'Medium', and 'High' ratings.

Sentiment EDA: With sentiment EDA, all "stop words" such as 'a', 'what', 'the', 'so', etc. will be removed from the written reviews, leaving words conducive to more accurate sentiment analysis. Stop words are defined as words that provide strictly low-level information to the text and are used to make the writing complete sentences [2]. They do not affect the sentiment analysis of any text, and thus it is standard practice for sentiment analysis to remove these words in order to have cleaner data and more concise results.

Sentiment Analysis (Low Ratings): After the data has been cleaned and explored, the nrc (National Resource Council emotion lexicon) algorithm can be run. This analysis will categorize each word as either positive or negative or some other "emotion" [3]. The algorithm places the words into emotions such as "anger", "anticipation", "disgust", "fear", "joy", "sadness", "trust", and "surprise" [2]. The nrc analysis will be useful in understanding how certain emotion 'buckets' can affect the overall numerical rating of a review. After the nrc algorithm has been run, afinn analysis will be performed which will rank the words in the reviews on a scale from -5 to 5 [3]. This model applies all words to a scale of positive and negative, based on the determined sentiment of the word as it is run against the data on which this model was originally created and trained. The final analysis performed for the low ratings will be bing analysis which will simply group the words into two categories, positive or negative [2].

Sentiment Analysis (Medium Ratings): The analysis for the medium ratings will follow the same process as the sentiment analysis for the low ratings. First, the nrc algorithm will be run on the reviews followed by afinn analysis and bing analysis.

Sentiment Analysis (High Ratings): The analysis for the high ratings will follow the same process as the sentiment analysis for the low and medium ratings. First, the nrc algorithm will be run on the reviews followed by afinn analysis and bing analysis.

Synthesis/Goal: After the completion of sentiment analysis for all three ratings groups, we will average the scores given by the afinn analysis within each rating group in order to quantify the data. This should be vital in answering our research question and evaluating our hypothesis – we would expect that the afinn score of the Low rating group will be the lowest or most negative, with the Medium rating group increasing and the High rating group having the highest average sentiment rating. Furthermore, we will plot the number of positive and negative words for each rating group by the bing analysis to understand the distribution of positive and negative words for each ranking group [4]. As for the nrc analysis, this will mainly help inform further research or analysis relating to our question. Once we understand the groupings of emotions that exist behind each group of text, we could take our analysis further and discover which specific words

are fuelling these emotions. While this is not completely out of the scope of our analysis, it would go beyond the basics of our research question and hypothesis and therefore may not be completely necessary.

References:

[1] D. Deshpande, “IMDB Movie Reviews 2021,” *Kaggle*, Dec, 2021. [Online]. Available: https://www.kaggle.com/datasets/darshan1504/imdb-movie-reviews-2021/code?datasetId=1379604. [Accessed Feb. 14, 2023].

[2] C. Khanna, “Text pre-processing: Stop words removal using different libraries,” *Towards Data Science,* para. 4, Feb 10, 2021. [Online]. Available: https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a. [Accessed Feb. 14, 2023].

[3] P. Sonkin, “Sentiment Analysis,” *Sentiment Analysis of 49 years of Warren Buffett’s Letters to Shareholders of Berkshire Hathaway,* para. 3, Jan. 19, 2021. [Online]. Available: https://bookdown.org/psonkin18/berkshire/sentiment.html. [Accessed Feb. 14, 2023]

[4] B. Boehmke, “Text Mining: Sentiment Analysis,” *UC Business Analytics R Programming Guide*, para. 12, 9 Jun, 2016. [Online]. Available: https://uc-r.github.io/sentiment_analysis.