B9DA109 Programming for Data Analysis: CA_TWO

IMDb Movie Reviews Sentiment Analysis

January 2025

Submitted by:
Anish Rao: 20066423
Lecturer: Muhammad Asad

Assignment Cover Sheet

Student Name and Number as per student card: Anish Rao - 20066423
Programme: Master of Science in Data Analytics (BMS09DNL)
Lecturer Name: Muhammad Asad
Module/Subject Title: B9DA108 – Programming for Data Analysis
Assignment Title: IMDb Movie Reviews Sentiment Analysis

By submitting this assignment, I am confirming that:

- **This assignment is all my own work.**
- **Any sources used have been referenced.**
- **I have followed the Generative AI instructions/ scale set out in the Assignment Brief.**
- **I have read the College rules regarding academic integrity in the QAH Part B Section 3, and the Generative AI Guidelines, and understand that penalties will be applied accordingly if work is found not to be my/our own.**
- **I understand that all work submitted may be code-matched report to show any similarities with other work.**

Index

1. Introduction

This projects aims to identify meaningful information by doing sentiment analysis on IMDb movie reviews. Sentiment analysis is an important area of Natural Language Processing, which involves analysing and determining data sentiment, like reviews/comments. Filmmakers, marketers and others need to understand audience attitude/feelings as it gives important information for improving quality of their content, audience engagement and general improvement. Analysing the sentiments also give more deeper understanding of audience preferences, which may help change/influence future choices.

Dataset Description

We used IMDb Movie Reviews Dataset for our analysis, containing a large-scale collection of movie reviews from IMDb (Maas et al., 2011), an online database providing vast information on movies, television shows, and celebrity profiles. The dataset has 50,000 user reviews that have labels with 2 sentiment classes positive or negative.

Objectives of the Analysis

1. Examining distribution of sentiments to check whether reviews are mostly positive or negative.
2. Determine and examine common words/terms used in positive versus negative reviews, giving insights into elements that affect audience perception.
3. Examine the connection between sentiment and review, to see if a specific sentiment affects review length.
4. Determine if sentiment intensity of short reviews is more extreme than of the longer reviews.
5. Analyse subjectivity and polarity to understand if reviews are more emotional or opinion-based.
6. Finding any trends or patterns in the most helpful or extreme reviews.

In order to get a better understanding of user-generated content, this research tries to give valuable insights about audience sentiment trends, review patterns and characteristics to get a deeper understanding of user-generated content.

2. Data Description

Dataset Details

The IMDb movie reviews dataset has 50,000 reviews with sentiment labels. It was compiled by researchers as a benchmark for analysing sentiments (Maas et al., 2011) and later made accessible on Kaggle.

- **Dataset Source:** IMDb Dataset (Lakshmi, 2020).
- **Dataset Size:** 50,000 user reviews.

Data Structure

Column Name	Description	Data Type
review	Textual movie reviews from users	Text (String)
sentiment	Binary classification labels of reviews, either "*positive*" or "*negative*".	Text (String)

Each entry in the dataset has individual movie reviews, with clearly labelled by sentiment, making it ideal for sentiment analysis.

Data Characteristics/Features

Initial overview:

- **Textual data**: Unstructured text data that required preprocessing.
- **Binary Sentiment Labels**: Values are categorical, either positive or negative.
- **Balanced Sentiments:** There are 25,000 reviews each for positive and negative reviews.
- **Duplicate Reviews:** 418 duplicates were removed, leaving the final dataset size to 49,582 unique reviews.
- **Varying Review Lengths**: The reviews varied in length from 4 to 2450 words

In order to prepare the dataset to analyse, it was cleaned and pre-processed. This includes converting all to lowercase, removing duplicates, html tags, punctuations and whitespace. The final dataset had 49,582 reviews.

Sentiment	Review count	Percentage
Positive	24884	50.19%
Negative	24698	49.81%

3. Method of Data Analysis

This sections describes the systematic approach of different analysis techniques used to fetch meaningful information/insights from the dataset.

Data Preprocessing

The following preprocessing steps were used to prepare the data for analysis:

- **Duplicate Removal**
  - There were 418 duplicate reviews that were removed, reducing dataset to 49,582 unique reviews.
- **Text Cleaning**
  - The text was changed to lowercase.
  - All HTML tags, punctuation and extra spaces were removed.
- **Removing stop-words**
  - Stop-words or non-informative words like “a”, “an”, “the”, “for”, etc. were removed using NLTK package.

These steps helped in cleaning the dataset for further analysis.

Data Analysis Techniques

The below analysis techniques were used:

1. **Sentiment Distribution Analysis**
   - Using a bar chart to display the number of positive and negative reviews.
2. **Word frequency and WordClouds**
   - Using WordClouds to determine the difference in vocabularies and the most frequent/common words in positive and negative reviews.
3. **Review Length Analysis**
   - Checking the word count in reviews to see if length and sentiments are related.
4. **Sentiment Intensity Analysis in Short vs. Long Reviews**
   - Used scatter plot to analyse the relationship between review length and polarity.
5. **Polarity and Subjectivity Distribution Analysis**
   - Generated histograms for clear visualisation of distribution of polarity and subjectivity.
6. **Most helpful or Extreme Reviews Analysis**
   - Identified reviews with extreme sentiments and longest reviews to recognise any patterns.

Tools and Implementation

- **Pandas and NumPy:** Used for data manipulation, getting statistics, preprocessing.
- **Matplotlib, Seaborn, Bokeh:** Used for creating visualisations.
- **WordCloud:** Used for visualisation of frequent words.
- **TextBlob:** Calculating polarity and subjectivity (Loria, 2018)
    - **Polarity:** it ranges between -1 (very negative) to +1 (very positive), to measure emotion.
    - **Subjectivity:** it ranges between 0 (objective) to 1 (subjective), to measure opinion.

4. Results

Below are the results of the Exploratory Data Analysis performed on the dataset.
Google Colab Link.

1. Sentiment Distribution

The dataset has 50.2% positive and 49.8% negative reviews. Meaning there is nearly equal distribution and guarantees unbiased analysis (Figure 1).

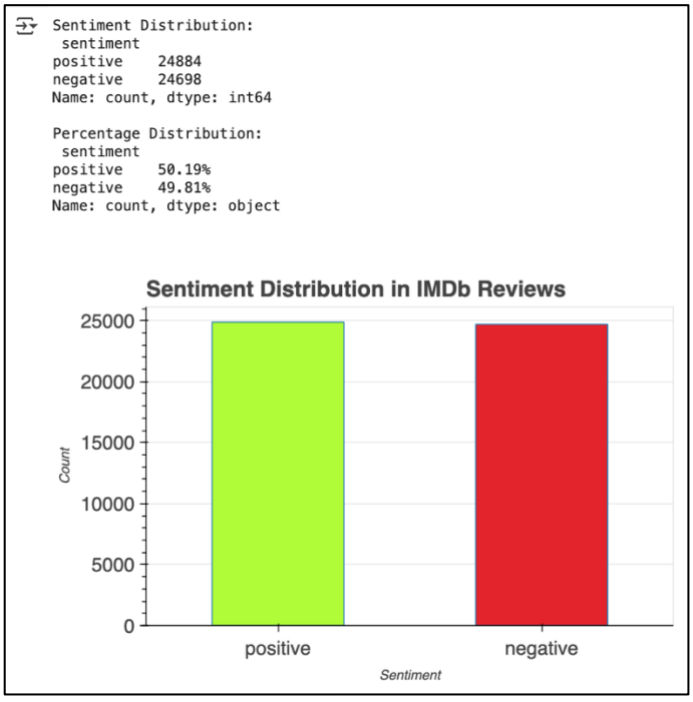

Figure 1

2. Most Common Words

Figure 3 shows the common words using a WordCloud in both positive and negative reviews. Figure 2 shows the counts of these words.

```
Total number of words in positive reviews: 5681185
Total number of words in negative reviews: 5543979

==================================================

Top 10 Most Common Words in Positive Reviews:
film: 39285
movie: 35836
one: 25621
like: 16999
good: 14286
great: 12570
story: 12338
see: 11814
time: 11724
well: 10933

Top 10 Most Common Words in Negative Reviews:
movie: 47010
film: 34651
one: 24364
like: 21509
even: 14761
good: 13997
bad: 13906
would: 13483
really: 12084
time: 11350
```

Figure 2

Figure 3

3. Review Length Analysis

Figure 5 shows a right-skewed histogram. We can say from this, most reviews are between 100 to 250 words.

```
Review Length Statistics:
count    49582.000000
mean       226.395950
std        167.728067
min          4.000000
25%        124.000000
50%        170.000000
75%        275.000000
max       2450.000000
Name: review_length, dtype: float64

================================================================================

              count        mean         std   min    25%    50%    75%     max
sentiment
negative   24698.0  224.470767  161.154345   4.0  125.0  171.0  272.0  1473.0
positive   24884.0  228.306743  173.989442  10.0  123.0  169.0  277.0  2450.0
```

Figure 4

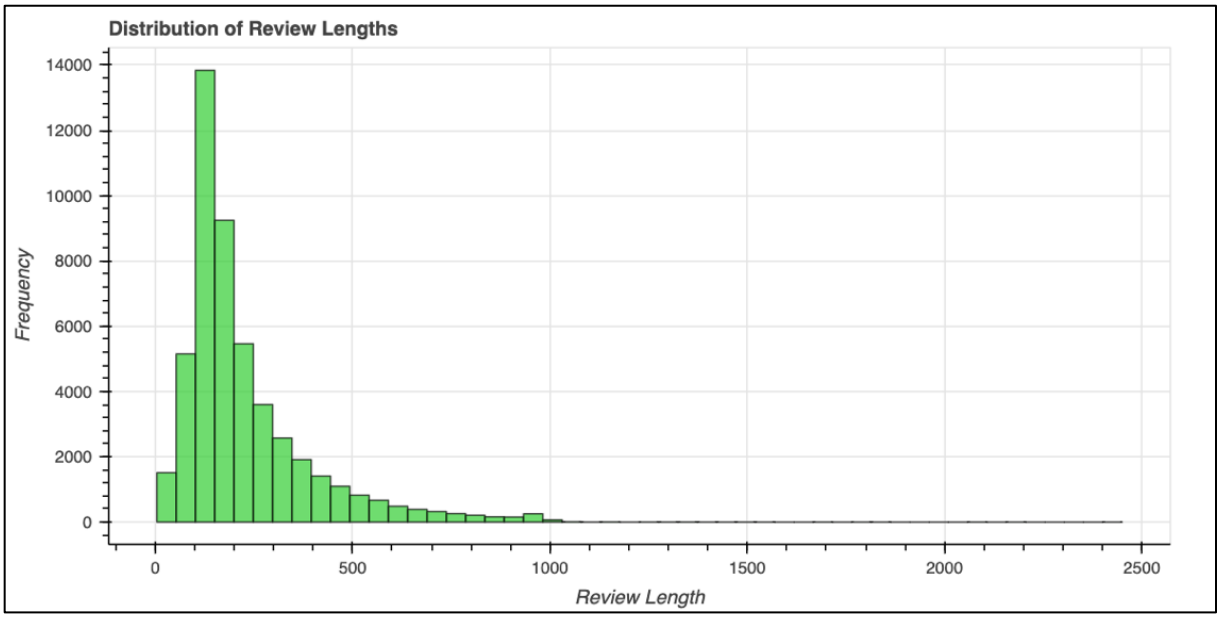

Figure 5

4. Sentiment Intensity vs. Review Length

Review length and sentiment polarity have a very weak negative correlation around -0.05 (Figure 6).

```
Sentiment Polarity Statistics:
count    49582.000000
mean         0.101630
std          0.161207
min         -1.000000
25%          0.003490
50%          0.102806
75%          0.200885
max          1.000000
Name: polarity, dtype: float64


Correlation between Review Length and Sentiment Polarity:
                review_length  polarity
review_length        1.000000 -0.049594
polarity            -0.049594  1.000000
```

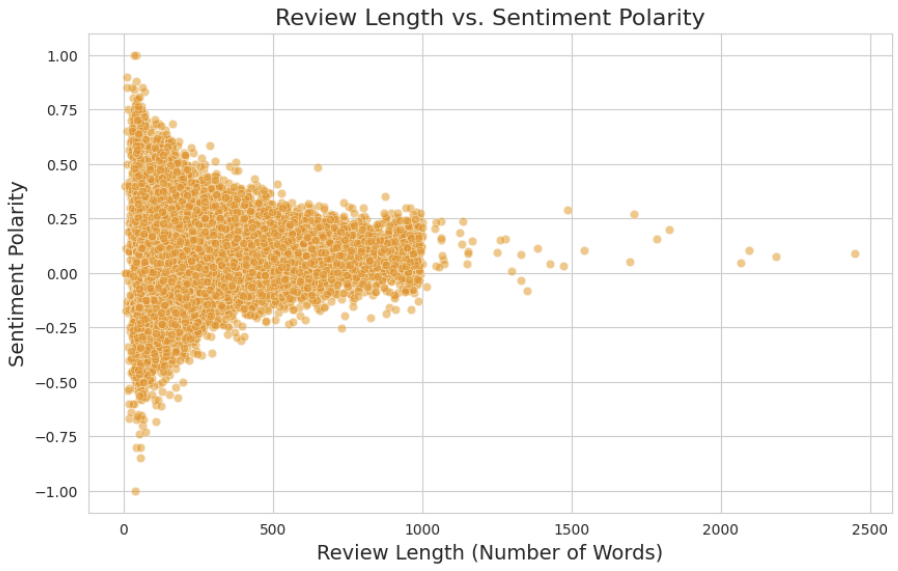

Figure 6

5. Polarity and Subjectivity Distributions

Most of the reviews are moderately subjective and positive polarity. (Figures 7 and 8)

```
Review Subjectivity Statistics:
count    49582.000000
mean         0.530600
std          0.092784
min          0.000000
25%          0.472421
50%          0.528820
75%          0.587449
max          1.000000
Name: subjectivity, dtype: float64


Polarity Distribution:
Highly Negative (-1 to -0.5): 57 reviews
Moderately Negative (-0.5 to 0): 11954 reviews
Neutral (0): 33 reviews
Moderately Positive (0 to 0.5): 37067 reviews
Highly Positive (0.5 to 1): 471 reviews

Subjectivity Distribution:
Highly Objective (0 to 0.3): 439 reviews
Moderately Objective (0.3 to 0.5): 17672 reviews
Moderately Subjective (0.5 to 0.7): 29653 reviews
Highly Subjective (0.7 to 1): 1818 reviews
```

Figure 7

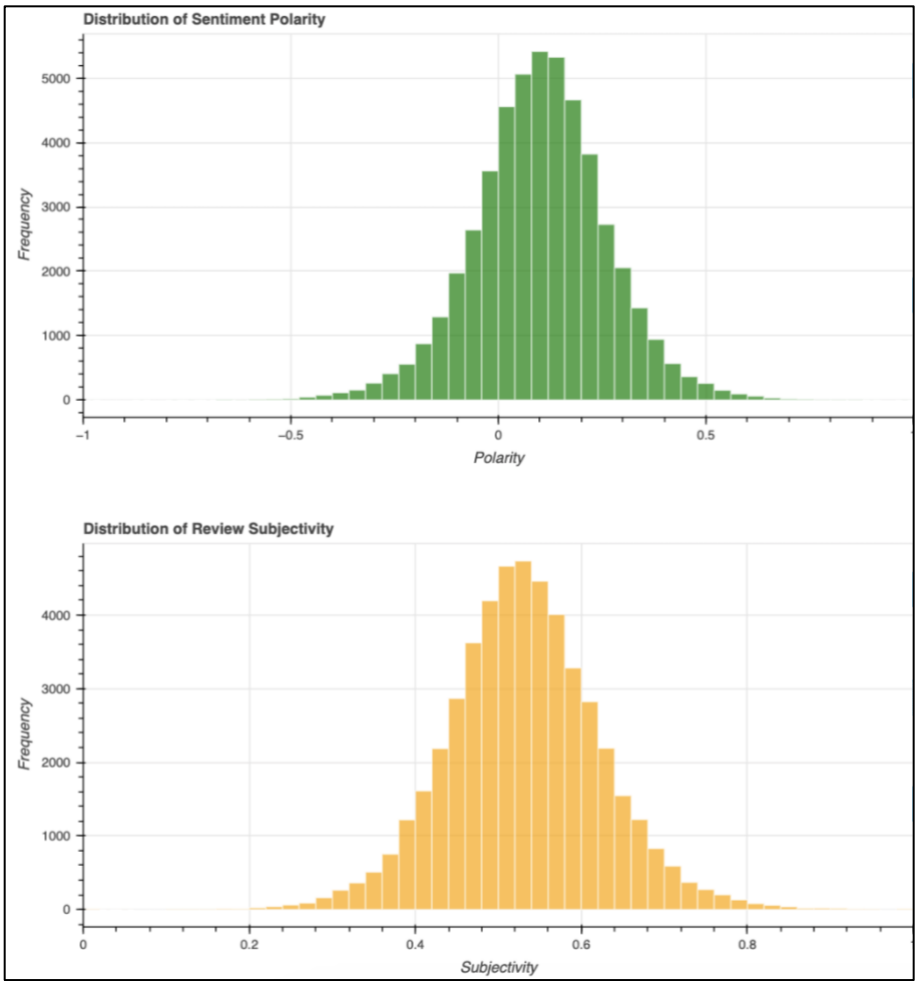

Figure 8

6. Extreme and Long Reviews

Figure 9 shows that extremely opinionated reviews are mostly short, while the longer reviews are more neutral.

Top 5 Reviews with Extreme Sentiment:

	review	cleaned_review	polarity	review_length
11973	This movie had me smiling from beginning to en...	this movie had me smiling from beginning to en...	1.0000	44
21532	Farley and Spade's best work ever. It's one of...	farley and spades best work ever its one of th...	1.0000	35
13865	This movie was horrible and corny. James Agee ...	this movie was horrible and corny james agee i...	-1.0000	38
18342	Brilliant and moving performances by Tom Court...	brilliant and moving performances by tom court...	0.9000	10
12484	I felt a great joy, after seeing this film, no...	i felt a great joy after seeing this film not ...	0.8775	41

==========

Top 5 Longest Reviews:

	review	cleaned_review	review_length	polarity
31309	Match 1: Tag Team Table Match Bubba Ray and Sp...	match 1 tag team table match bubba ray and spi...	2450	0.089456
40247	There's a sign on The Lost Highway that says:<...	theres a sign on the lost highway that saysmaj...	2186	0.075269
31265	Back in the mid/late 80s, an OAV anime by titl...	back in the midlate 80s an oav anime by title ...	2094	0.102263
31072	(Some spoilers included:) Although,...	some spoilers includedalthough many commentato...	2068	0.047715
12622	Titanic directed by James Cameron presents a f...	titanic directed by james cameron presents a f...	1827	0.199990

Figure 9

5. Discussion

Below are interpretations and key insights derived from the results of the visualisations.

1. Sentiment Distribution

There is an equal distribution of reviews having 50.2% positive and 49.8% negative reviews. Because of this equality the data is not biased making it perfect for unbiased sentiment analysis.

2. Most Common Words

The majority of the terms in positive reviews are "great", "best", “good” and “fun”, but words in negative reviews are “bad”, “awful”, “worst” and “boring”. Both sentiments mention “movie” and “film” which means they both focus on the movie experience and have similar tones but very different contexts.

3. Review Length Analysis

This histogram of review lengths is right-skewed. Majority of the reviews are between 100 to 250 words. Then average word count of negative and positive reviews is almost same, with positive having 228 and negative having 224 reviews. There are only 41 reviews with more than 1,000 words, this suggests that review length is not impacted strongly by sentiment.

4. Sentiment Intensity vs. Review Length

The scatter plot shows that short reviews have high polarities, whereas longer reviews all cluster around neutral polarity. The weak negative correlation (approximately -0.05) suggests longer reviews are more focused on the details and shorter reviews are emotional.

5. Polarity and Subjectivity Distributions

The polarity histogram displays most reviews are mildly positive (0 to 0.5), and extreme sentiments are rare. The subjectivity graph displays most reviews are moderately subjective (0.5 to 0.7), meaning reviews are not extreme and highly objective/subjective reviews are rare.

6. Extreme and Long Reviews

Most extreme sentiment reviews are around 10 to 50 words which confirms that shorter reviews have stronger emotion. However the longest reviews have almost neutral polarity meaning they focus more on the details of the movie.

Limitations and Challenges

The main drawback is relying on TextBlob as it might misread sarcasm or irony in some reviews. Also there is very little text preprocessing done which means there is some leftover noise or ambiguity in the data, affecting the accuracy slightly.
Additionally there was a lack of context information like reviewer demographics, genre, move star ratings which could have provided deeper insights. Without user upvotes, getting the actual helpfulness of reviews was not possible.

6. Conclusion

This sentiment analysis of IMDb movie reviews successfully shows meaningful patterns in user opinions, usage of vocabulary, sentiment distribution, and review characteristics. One of the advantages of the dataset is that it has almost same distribution, meaning 50% were positive and 50% were negative reviews. This gave us a strong foundation for comparisons.

The analysis showed that there were few words that were common in both positive and negative reviews, but there were visible differences in the vocabulary. Positive reviews often praised elements like storytelling, acting, or emotional impact using favourable words, on the other hand negative reviews focused more on expressing dissatisfaction and criticism with stronger, negative language. The length of a review did not strongly affect whether it was positive or negative. However, shorter reviews often showed stronger emotions, both very positive and very negative. While longer reviews had a balanced tone and were usually more detailed. TextBlob's polarity and subjectivity analysis showed that extreme sentiments were uncommon. Most users preferred subtle, moderately positive expressions, adding objective details with personal opinions. This shows a general trend among users an overall tendency among the users to give balanced feedback instead of extreme assessments.

This analysis effectively achieved the main goal even with drawbacks like using a simple sentiment tool (TextBlob), lack of review helpfulness metrics like user votes and less text preprocessing. It still was able to provide insightful information from user reviews.

For future work, using advanced Natural Language Processing methods like transformer-based models (e.g., BERT) could help in increasing detection of sentiments with more context. Topic modelling can also help us find any hidden themes across different genres, and adding more review metadata like helpfulness votes (user voting on reviews) could also help verify assumptions about what makes a review impactful. We could also compare IMDb with other platforms for reviews and check if these insights are similar across domains. Overall, this study has given a strong foundation for further analysis of user-generated content in the entertainment media.

7. References

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C. (2011) *Learning word vectors for sentiment analysis*, *ACL Anthology*. Available at: https://aclanthology.org/P11-1015/ (Accessed: 15 March 2025).

N, L. (2019) *IMDB dataset of 50K movie reviews*, *Kaggle*. Available at: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data (Accessed: 15 March 2025).

Loria, S. (no date) *Simplified text processing*, *TextBlob*. Available at: https://textblob.readthedocs.io/en/dev/ (Accessed: 17 March 2025).