

Movie Reviews and their relation to the ratings on IMDb

Megan Hardy s191198
Hideaki Fukuyama s221591

January 14, 2026

Outline

Research Question

Database

Data Cleaning & Management

Visualizations

Modeling

Conclusion

Research Question

- ▶ Last pitch: "Predicting Movie Review Sentiment from IMDb", we wanted to determine if a movie review is positive or negative based on the text.
- ▶ New research question: "Can IMDb ratings be predicted from reviews?"
- ▶ Is the relationship improved after adding other components such as metadata and reviewer information?

Sources of the Database

- ▶ IMDb Metadata Up-to-Date: All informations from the website available (genre, release year, title, id...), using `title.ratings.tsv` and `title.basics.tsv`.
- ▶ IMDb Review Dataset (Kaggle): `.json` files with titles associated with reviews, concatenated into a single Polars Dataframe
- ▶ The Review Dataset is merged with the Movie Metadata using the normalized title as the key.
- ▶ We are taking a sample of 500 000 reviews out of the millions of review that we have in the original database.

Overview of the Cleaned Dataset

shapes: (5, 17)

review_id	reviewer	movie	rating	review_summary	review_date	review_detail	clean_title	title_norm	tconst	titleType	primaryTitle	startYear	runtimeMinutes
str	str	str	str	str	str	str	str	str	str	str	str	str	str
"rw2120619"	"lmsnowhte"	"Buddy (1997)"	null	"Help! to find Gorilla Movie Na..."	"1 September 2009"	"Does anyone remember the name ..."	"Buddy"	"buddy"	"tt8819170"	"movie"	"Buddy"	"2018"	"86"
"rw4121044"	"docwebb-65066"	"Amateur (2018)"	"6"	"Somewhat believable"	"8 April 2018"	"Pretty good throughout the mov..."	"Amateur"	"amateur"	"tt18568060"	"movie"	"Amateur"	"2022"	"75"
"rw1678078"	"ccthemoviemanager-1"	"Assault on Precinct 13 (1976)"	"1"	"Another Unbelievably Overrated..."	"21 June 2007"	"Why is this film such a critic..."	"Assault on Precinct 13"	"assault on precinct 13"	"tt0074156"	"movie"	"Assault on Precinct 13"	"1976"	"91"
"rw1757206"	"yeah_sure"	"Invincible (2001)"	"8"	"Simple, interesting, and live..."	"2 November 2007"	"I bought this DVD purely on a ..."	"Invincible"	"invincible"	"tt0245171"	"movie"	"Invincible"	"2001"	"133"
"rw4236164"	"aswilliams40"	"Trust (2018)"	"9"	"Where Did TRUST go?"	"13 July 2018"	"I was watching it & now I can'..."	"Trust"	"trust"	"tt0828461"	"movie"	"Trust"	"2006"	"86"

Data Management

We are going to apply a text-cleaning function to remove the HTML tags, emojis and text is converted to lowercase. Removing punctuation except for exclamation and question marks.

We are going to create variables that influences the IMDb ratings:

- ▶ Review length (Engagement and emphasis)
- ▶ Vocabulary Richness (Linguistic diversity)
- ▶ Sentiment Analysis (VADER to capture sentiment, Textblob to capture subjectivity)
- ▶ Movie age
- ▶ Reviewer features (Number of reviews, Average reviewer rating, Reviewer Bias, Heavy Reviewers)
- ▶ Genre categorization

Visualizations

For the descriptive analysis, we have 9 modelizations:

- ▶ **Distribution of IMDb rating**
- ▶ **Relation sentiment (Vader) vs rating**
- ▶ **Relation between review length vs rating**
- ▶ **Relation between review emphasis vs rating**
- ▶ **Relation between review subjectivity vs rating**
- ▶ **Relation between review group vs rating**
- ▶ **Relation between review bias vs rating**
- ▶ **Mean rating per genre**
- ▶ **Heatmap correlation**

Five of the most relevant visualizations will be presented

Distribution of IMDb rating

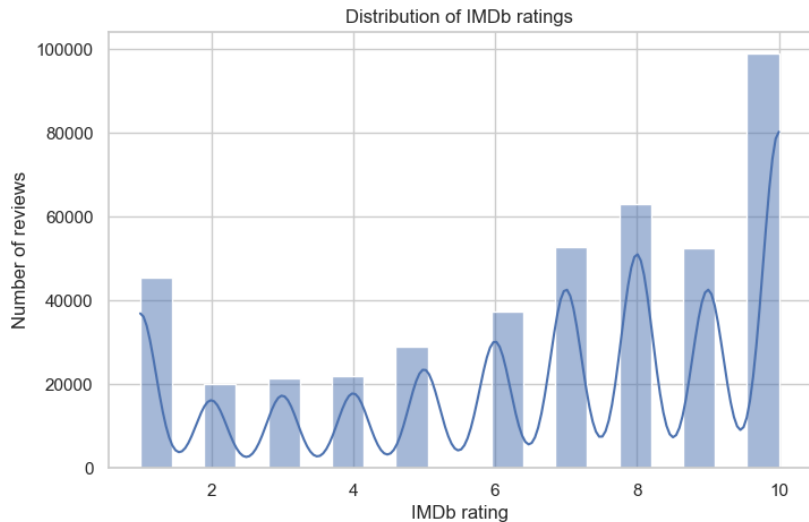


Figure 1: Distribution of IMDb rating

Relation Sentiment (VADER) vs Rating



Figure 2: Relation between VADER and IMDb ratings

Relation Reviewer Group vs Rating

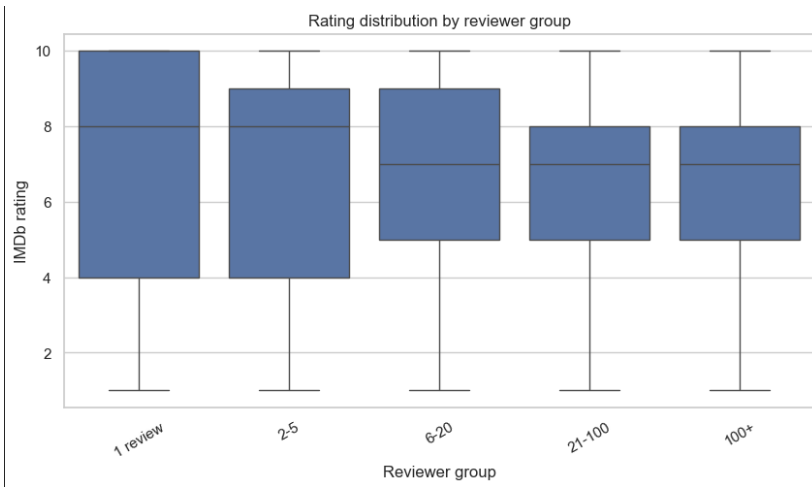


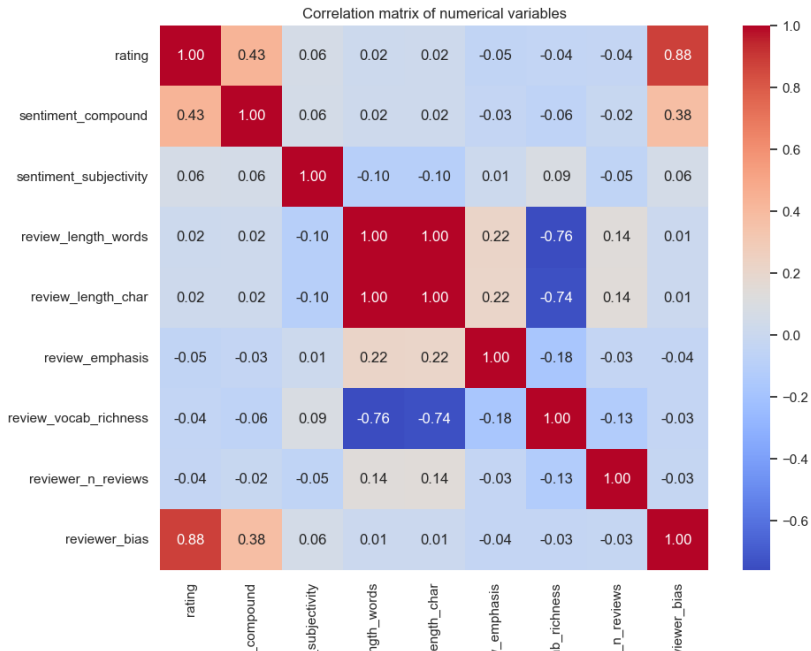
Figure 3: Relation Between Reviewer Groups and IMDb ratings

Relation Reviewer Biases vs Rating



Figure 4: Relation Between Reviewer Biases and IMDb ratings

Correlation Heatmap of all numerical variables



Linear Regression Model

We have decided to do three different models to test three hypotheses:

- ▶ Linear Regression (NLP): Only using two main variables, with sentiment_compound, sentiment_subjectivity and the interaction of those terms (subjectivityxsentiment)
- ▶ Linear Regression (NLP + Metadata): We are adding the variables movie_age, average_rating, numVotes and primary_genre to test if the movie context helps prediction.
- ▶ Linear Regression (NLP + Metadata + Reviewer): We are adding the final variables: reviewer_bias, reviewer_n_reviews and reviewer_group in order to test the impact of the reviewer behaviour.

Linear Regression Model: Results

	Model	RMSE	R2
0	Simple Linear Regression (1)	2.673177	0.196437
1	Model with metadata (2)	2.642792	0.212481
2	Final model with metadata and review features (3)	1.408741	0.776232

Figure 6: Comparative table between the 3 models

Linear Regression Model: Coefficients

```
primary_genre_Music      -0.158302
primary_genre_Comedy     -0.113680
primary_genre_Action     -0.084706
primary_genre_Adventure  -0.080211
primary_genre_Adult      -0.076383
primary_genre_Musical    -0.075423
primary_genre_Documentary -0.055525
primary_genre_Film-Noir  -0.055253
primary_genre_Fantasy    -0.048215
primary_genre_Horror     -0.045116
primary_genre_Romance    -0.040685
primary_genre_Mystery    -0.036947
primary_genre_Animation  -0.034765
primary_genre_Drama      -0.030583
primary_genre_Unknown    -0.022988
reviewer_group_6-20     -0.009158
reviewer_group_21-100   -0.005003
sentiment_subjectivity  -0.003969
reviewer_n_reviews      -0.002360
primary_genre_Reality-TV 0.000414
reviewer_group_1 review 0.001728
reviewer_group_2-5      0.002464
numVotes                0.009596
reviewer_group_100+     0.009969
primary_genre_Thriller   0.012895
...
primary_genre_Sport      0.149178
subjectivity_x_sentiment 0.292908
primary_genre_War        0.369625
reviewer_bias            2.453946
dtype: float64
```

Figure 7: Coefficients of the Linear Regression

TF-IDF & Linear Regression

Configuration:

- ▶ Max features: 5,000 features
- ▶ Minimum document frequency: 5 (words appearing in fewer than 5 reviews excluded)
- ▶ Removing english stopwords

Metric	Value
RMSE	1.975
R ²	0.561

Table 1: TF-IDF Performance

```
Top positive words influencing rating:
best      4.688199
excellent  4.471692
great     4.258627
loved     4.164188
amazing   4.136191
1010      4.038855
perfect   4.006590
brilliant 3.977481
complaining 3.860828
awesome   3.761828
mustsee   3.665087
favorite  3.577397
refreshing 3.565469
greatest 3.505093
superb    3.444820
funniest  3.403276
masterpiece 3.380089
phenomenal 3.306235
complain  3.298656
910       3.275379
dtype: float64

Top negative words influencing rating:
worst     -8.598255
...
disgusting -4.750638
downhill  -4.739767
propaganda -4.659573
dtype: float64
```

Figure 8: TF-IDF Coefficients

Random Forest

Configuration:

- ▶ Number of estimators: 200 trees

Metric	Value
RMSE	1.348
R^2	0.795

Table 2: Random Forest Model Performance

```
reviewer_bias      0.785908
sentiment_compound 0.045906
numVotes           0.033874
subjectivity_x_sentiment 0.029037
averageRating      0.027116
sentiment_subjectivity 0.027087
reviewer_n_reviews 0.025800
movie_age          0.025274
dtype: float64
```

Figure 9: Random Forest Feature Importance

Conclusion

- ▶ Prediction depends on both review content and reviewer behavior.
- ▶ Reviewer bias is the strongest predictor of ratings.
- ▶ TF-IDF shows text matters, but alone it's insufficient.
- ▶ Random Forest achieves the best predictive accuracy, capturing non-linear relationships.
- ▶ Linear Regression provides interpretability and highlights the role of reviewer bias + sentiment \times subjectivity.