Topic Modeling Drivers of Movie Review Sentiments

Parisa Mershon, Joseph Lobowicz

So What?

In this project, we apply latent Dirichlet allocation topic modeling to a corpus of IMDB movie reviews to uncover the themes that drive positive and negative sentiment. Next, we select an optimal number of topics via perplexity, log-likelihood, and coherence criteria. We then interpret the resulting topics, compare their prevalence across sentiment groups, and build a gradient boosted tree classifier to predict sentiment from topic proportions.

Ultimately, we want to identify the latent themes (topics) that distinguish positive (“pos”) vs negative (“neg”) reviews in the IMDB dataset. By uncovering these topics, we can understand what aspects of a film drive audience praise or criticism, and then evaluate whether these topic features can predict sentiment.

Exploratory Data Analysis

Before diving into topic modeling, we explore the raw distribution of ratings and review lengths in our IMDB dataset to understand potential biases and how they may influence downstream analyses.

To begin, we tally the counts of each rating by sentiment label and plot them (see Figure 1).

- Rating 1 dominates the negative reviews with 5,022 observations, while Ratings 2–4 each contribute roughly 2,300–2,600 negatives.
- On the positive side, Rating 10 peaks at 2,971, with Ratings 7–9 trailing between 1,378 and 1,731.

This U-shaped curve indicates that reviewers seldom choose middle scores, opting instead for strong approval or disapproval. For topic modeling later, this means that thematic signals may be most pronounced at the extremes. Next, we inspect how review length varies across star ratings by plotting word-count distributions (see Figure 2).

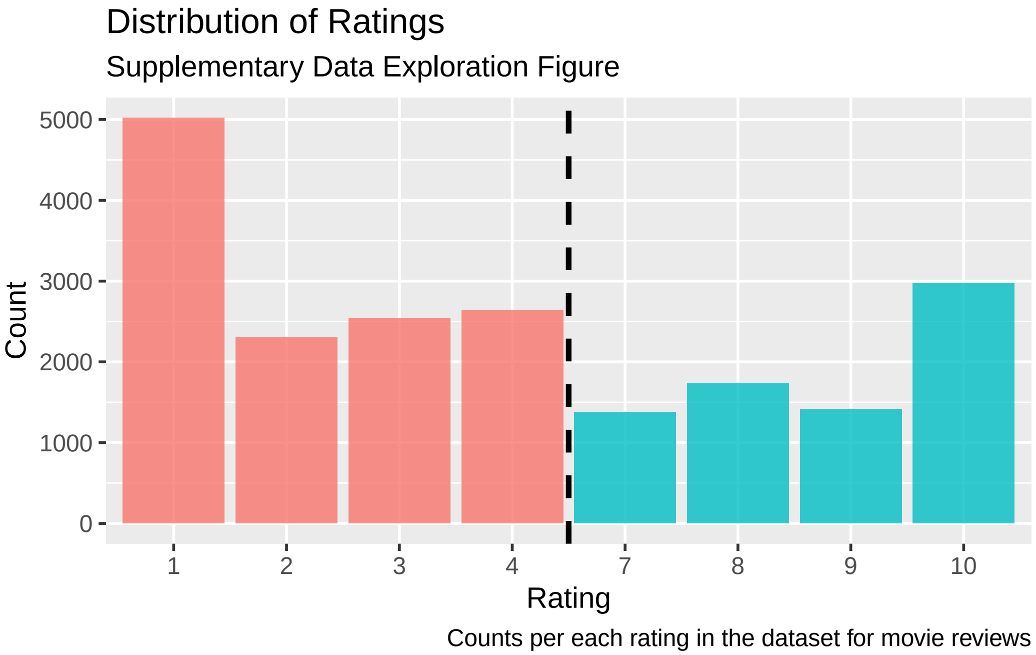


Figure 1

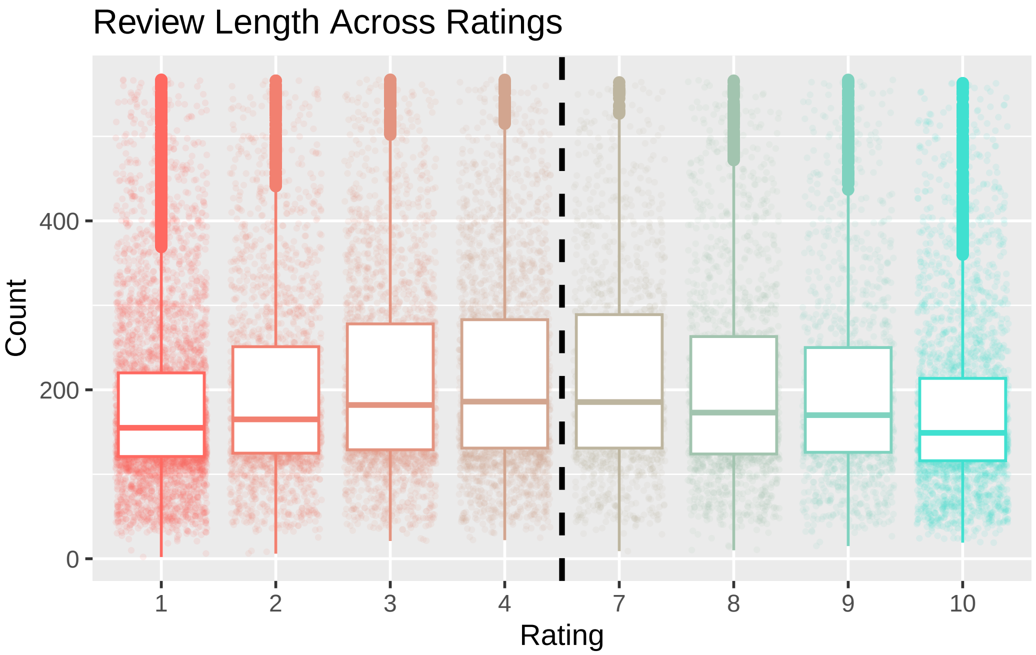


Figure 2

- Ratings 4 and 7 show the highest median word counts, suggesting that “middle” reviews tend to be more wordy.
- Extremes (1 and 10) yield the shortest comments, implying that strongly felt opinions may be conveyed more briefly.

Longer, midrange reviews may blend positive and negative terms, potentially creating topic mixtures that challenge clean separation. In contrast, shorter extreme reviews might contain only a few key sentiment words, skewing topic proportions toward dominant themes.

Clean Up

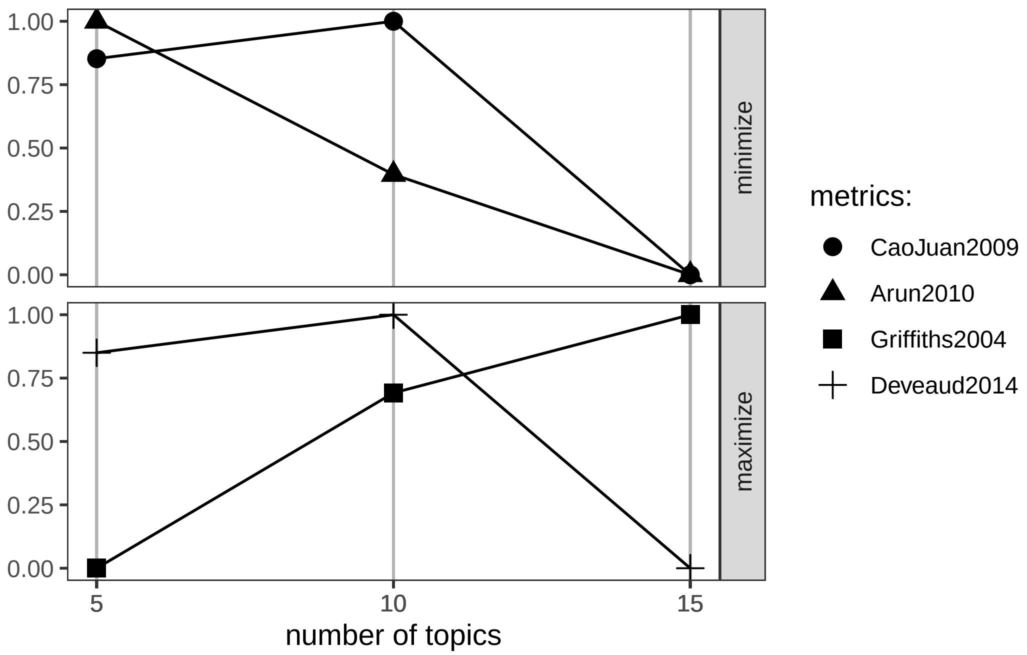
The dataset comprises 50,000 labeled reviews evenly split between training and testing. In order to save time running models, we will use only the rows with `split = train`. This subset contains 25,000 reviews. We remove all `<br /><br />` tags to avoid spurious tokens, then tokenize and drop stop words. The resulting tokens feed into a sparse document–term matrix for the training subset. This preprocessing ensures our topic models focus on terms without HTML noise.

Number of Topics

Having characterized basic corpus properties, we apply LDA to the training split’s document–term matrix. To choose the number of topics, we fit LDA models for `k = 5, 10, 15`, and compute both perplexity/log-likelihood and four coherence metrics as shown below.

```
# A tibble: 3 x 3
      k perplexity log_likelihood
  <dbl>      <dbl>          <dbl>
1     5      6386.     -17673543.
2    10      5915.     -16916950.
3    15      5624.     -16529228.
```

```
Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
of ggplot2 3.3.4.
i The deprecated feature was likely used in the ldatuning package.
  Please report the issue at <https://github.com/nikita-moor/ldatuning/issues>.
```



Perplexity steadily decreases while log-likelihood increases with k, but coherence peaks at ten topics balances the maximization and minimization of factors the best. So we select the optimal value of `k` to be 10.

Final Model & Top Terms Per Topic

Based on these diagnostics, we fix `k = 10` and refit the final LDA model. Its top ten terms per topic, as shown in Figure 3, form coherent clusters that we can label descriptively.

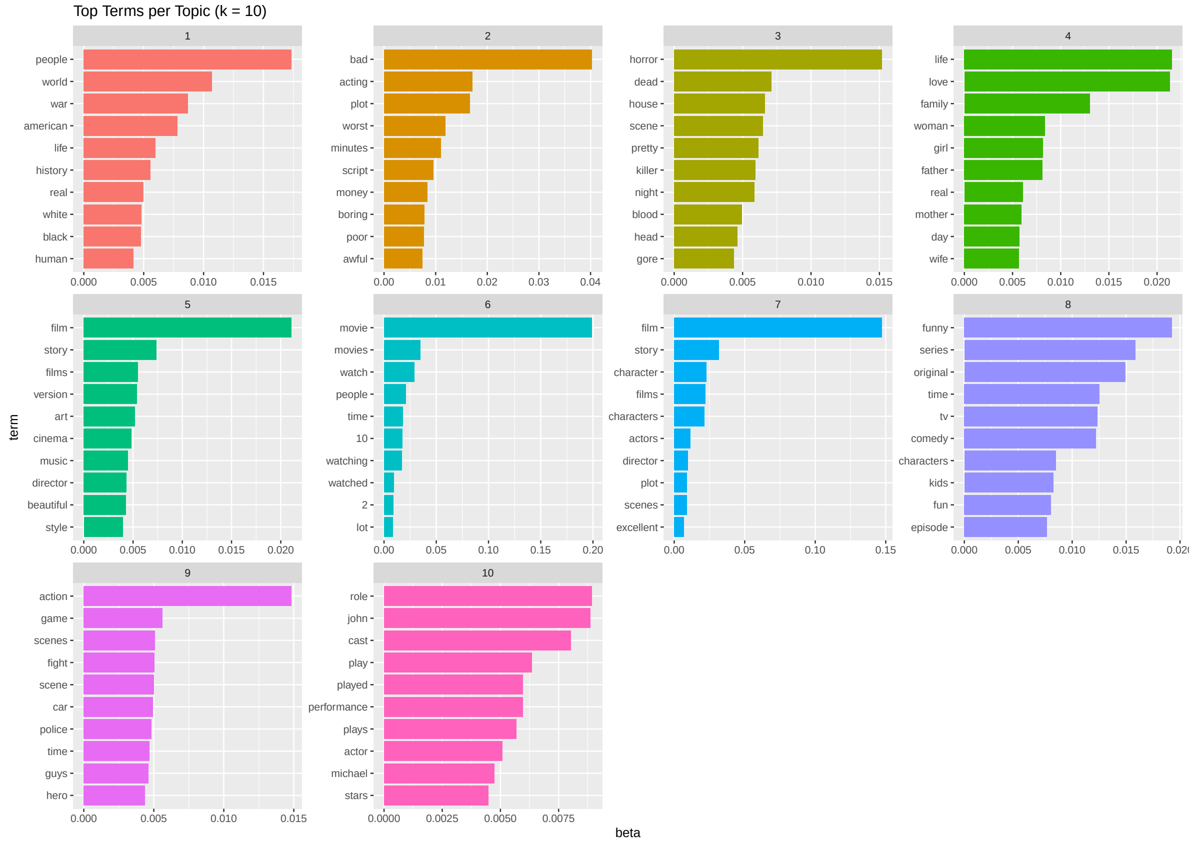


Figure 3

Some topics stick out as standout themes: Topic 2 collects strongly negative performance adjectives (“bad”, “boring”, “awful”), while Topic 3 captures horror-genre language (“horror”, “killer”, “gore”). Topics 5 and 7 both reference “film” and “director” but diverge into aesthetic critique (“beautiful”, “style”) versus character analysis (“character”, “actors”). These coherent term sets allow clear thematic labels. Consider,

- Topic 1: Social & Human Issues (world, war, history, life, human)
- Topic 2: Performance Critique (bad, acting, boring, script, plot, poor)
- Topic 3: Horror (dead, killer, night, blood, gore)
- Topic 4: Family/Relationships (life, mother, wife, love, family, girl)
- Topic 5: Art & Cinematography
- Topic 8: TV Series
- Topic 9: Action
- Topic 10: Casting & Roles

Topic Proportions

For each topic, compute its average prevalence in positive vs. negative reviews to see which themes drive praise and criticism.

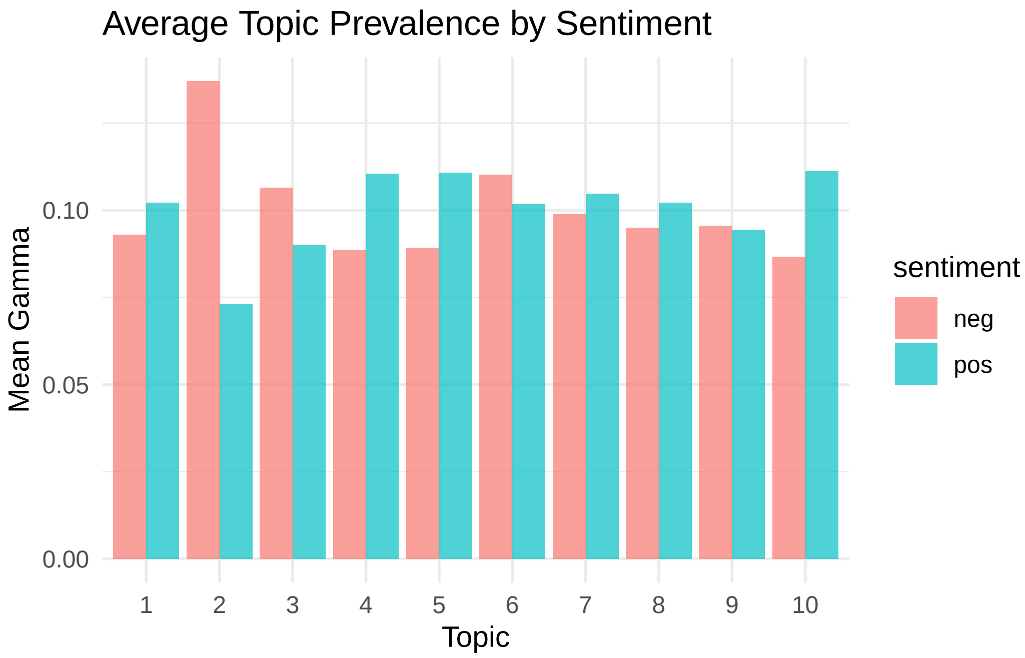


Figure 4

From Figure 4, we can see that that Topic 2 (Performance Critique) and Topic 3 (Horror) dominate negative reviews, whereas Topic 10 (Casting & Roles) and Topic 5 (Art & Cinematography) prevail in positive reviews. These themes are what reviewers mostly criticize for either sentiment, and could highlight which themes underlie praise and criticism.

Predictive Modeling with Topic Features (xgboost)

Our goal for this section is to use each review's topic proportions as features to predict whether it's a positive review (otherwise, a negative one). We trained a gradient boosted tree model because it can capture non-linear interactions among topics and is robust to multicollinearity in our topic-proportion features.

NB: We convert sentiment to factor 1 if sentiment is positive, and 0 if sentiment is negative.

```
  Feature        Gain      Cover  Frequency
   <char>       <num>      <num>      <num>
```

```
 1:  topic_2 0.79161319 0.45406193 0.18477092
 2:  topic_3 0.04616905 0.11151428 0.11558041
 3:  topic_1 0.02674982 0.07794018 0.10891231
 4:  topic_6 0.02451609 0.06074668 0.09851581
 5:  topic_5 0.02421975 0.07048902 0.10396501
 6:  topic_9 0.02203018 0.05945114 0.09722521
 7:  topic_4 0.01779899 0.04405084 0.07772281
 8:  topic_7 0.01748579 0.04617400 0.08073421
 9: topic_10 0.01617477 0.04495763 0.07177171
10:  topic_8 0.01324235 0.03061431 0.06080161
```

From the results above, we see that Topic 2 ("Performance Critique") dominates with nearly 80% of total gain, indicating that words like *bad*, *acting*, *plot*, *boring*, *script* are by far the strongest predictors of a review's sentiment. The next most important topics are Topic 3 (Horror) and Topic 1 (Social & Human Issues), accounting for under 5% of gain.

```
          Truth
Prediction    0    1
         0 1945  408
         1  555 2092
```

From the confusion matrix, 76.64% are true negatives, and 83.52% are true positives. The model slightly under-predicts negative reviews (584 false positives) and under-predicts positive reviews (412 false negatives).

```
# A tibble: 1 x 3
  .metric  .estimator .estimate
  <chr>    <chr>          <dbl>
1 accuracy binary         0.807
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.877
```

An AUC of 0.873 tells us that our topic-based GBT model has strong discriminative power. If we randomly pick one positive review and one negative review, there's about an 87% chance the model will assign a higher "positive" probability to the actual positive one. Our topics capture most of the signal needed to tell praise from criticism. Combined with ~80% accuracy, the AUC shows the model is confident and correct across the full spectrum of review.

Using non-linear GBTs over topic proportions yields a good sentiment predictor that underlines that “Performance Critique” (Topic 2) and the other top topics truly drive the positive/negative split.

Conclusion

In summary, our unsupervised topic model finds coherent themes that characterize and predict sentiment in movie reviews. The prominence of performance critique in negative reviews and casting/artistic topics in positive ones provides actionable insight into reviewers’ priorities. By coupling interpretable topics with a robust classifier, we reveal both the language patterns that drive audience reactions and the feasibility of sentiment prediction based solely on thematic content.