I, \_\_Xingyu Chen\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, understand that this is an individual take-home exam. I understand and agree that I will not work with any other humans or AI mechanisms to complete this exam. I agree on my honor to only do my own work, not to share or request any assistance on this exam, and to use only the resources permitted which include Dr. Gates code, sites, and resources, my own code, web resources that are not AI based, anything on Canvas, and any books/articles, etc. I understand and agree that sharing on an exam is unethical and unfair to my fellow classmates. I also agree that if sharing occurs, I will accept a 0 grade. I also agree to submit my Exam before the deadline. The absolute last moment to submit is May 5 at 12 noon MT. I understand that it is impossible to submit after this even by one second 😊

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Xingyu Chen \_\_\_\_\_April 26 2023\_\_\_\_\_\_Date

Introduction

The movie industry heavily relies on audience reception and feedback to the current movie so they can produce successful and profitable movies in the future. The sentiment analysis of movie reviews plays an important role for reveal the perception of the audience towards a movie, which can eventually impact its commercial success. Therefore, stakeholders in the movie industry, including producers, director, and marketing teams, are interested in understanding the sentiment of movie reviews to make better decisions about movie production and marketing strategies.

We have a dataset of movie reviews, consisting of both positive and negative reviews. Using text mining techniques to analyze this data, and gain insights into the factors that influence the sentiment of reviews. By identifying common themes and topics in positive and negative reviews, they hope to gain a better understanding of the key drivers of audience perception.

In addition to understanding the sentiment of existing reviews, the stakeholders are also interested in using text mining to predict the sentiment of future reviews. This can help them to tailor their marketing and distribution strategies to the target audience and maximize the success of their movies.

Data Cleaning

Word cloud for raw dataframe which label is Negative

A picture containing text, newspaper

Description automatically generated

Word cloud for raw dataframe which label is Positive

Text

Description automatically generated with medium confidence

Word cloud for cleaned dataframe which label is Positive

Text

Description automatically generated

Word cloud for cleaned dataframe which label is Negative

A picture containing text, newspaper

Description automatically generated

Raw Data:

A screenshot of a computer

Description automatically generated with low confidence

Clean Data:

Text

Description automatically generated

DataFrame with lemming:

Graphical user interface

Description automatically generated

DataFrame with 30 features:

A screen shot of a computer

Description automatically generated with low confidence

Part 3: Latent Dirichlet Allocation

Timeline

Description automatically generatedChart, bubble chart

Description automatically generated

Chart

Description automatically generated

Most of movie talks about the scene of the movie. The more detailed the scene, the more likely it is a positive review. On the other hand, the ‘love’, ‘like’, ‘good’, ‘character’ contributes to positive reviews. Meanwhile, it is hard to see what words contribute to negative reviews, need to use other methods to find association between them.

Part 4: Association Rule Mining

Chart, histogram

Description automatically generated

Top 10 rules for support, confidence, lift

Table

Description automatically generated

Top 10 rules for pos -> and neg ->

Table

Description automatically generatedTable

Description automatically generated with medium confidence

One network

Chart, bubble chart

Description automatically generated

From the graph and association rules we can see that positive reviews always talks about performance and character. They often comment about interesting characters and scenes. The ‘comedy’ often connects with ‘funny’ in the positive reviews. Both negative and positive reviews talk about ‘movies’, ‘story’, and ‘characters’ so it is hard to determine sentiment just from these three words. Negative reviews mention a lot of ‘really’, along with ‘effects’ and ‘special’.

Part 5: Clustering

Chart, scatter chart

Description automatically generated

Top 5 features that have the highest absolute difference between the two centroids:

- Feature 5: films

- Feature 21: nthis

- Feature 9: little

- Feature 20: nthe

- Feature 18: nin

From the k-mean visualization, we can clearly see that positive reviews and negative reviews are sparse in the graph, so it is hard to form two different clusters to divide these 40 testing reviews (20 actual positive and 20 actual negative). The main reason that we already saw that in the LDA method because ‘movie’, ‘character’, ‘film’ is all use frequently and occur in positive and negative reviews. We can see the margin from calculating the Euclidean distance between the test vectors and the centroids and the top features are films, this, little, the, and in.

Part 6: Naïve Bayes, Decision Trees, and SVMs

NB:

Chart

Description automatically generated

DT:

Df1



Df2



Df3



The confusion matrix df1 is:

[[292 28]

[224 56]]

The confusion matrix df2 is:

[[301 18]

[228 53]]

The confusion matrix df3 is:

[[200 225]

[ 96 279]]

SVM:

Linear:

The confusion matrix df1 is:

[[266 54]

[ 84 196]]

The confusion matrix df2 is:

[[254 65]

[ 71 210]]

The confusion matrix df3 is:

[[287 138]

[160 215]]

RBF:

The confusion matrix df1 is:

[[320 0]

[280 0]]

The confusion matrix df2 is:

[[319 0]

[281 0]]

The confusion matrix df3 is:

[[320 105]

[210 165]]

POLY:

The confusion matrix df1 is:

[[320 0]

[280 0]]

The confusion matrix df2 is:

[[319 0]

[281 0]]

The confusion matrix df3 is:

[[262 163]

[201 174]]

Part 7: Conclusions

Take-home messages:

1. Positive movie reviews often contain keywords ‘love’, ‘like’, and ‘good’
2. Negative movie reviews often contain keywords ‘character’, ‘effect’, and ‘scene’
3. Both negative and positive reviews talk about ‘movies’, ‘story’, and ‘film’ so it is hard to determine sentiment from these features.
4. There is high frequency of ‘nis’, ‘nthere’, ‘nthe’. These weired words occurs due to bad raw data and when merge ‘\n’ and ‘There’, it will result in ‘nthere’ after data cleaning.
5. Due to evenly distributed dataset (half of them are negative, others are positive), we do not need down sampling or up sampling the dataset and data model provide better prediction result.
6. The length of movie reviews is around 2500 – 3500 characters length or in other words, 120 - 150 words length.
7. It is better to have a Neutral label or numerical values from 1 to 5 for better sentiment analysis
8. The more complex the model, the better performance it will provide but less interpretation
9. The larger of parameter put into the model, more time and memory it will consume but it does not guarantee better result
10. This paper showcases the significance of text mining in gaining valuable insights into movie audience opinions.

|  |  |  |  |
| --- | --- | --- | --- |
|  | DF1 (CV) | DF2 (CV lemm) | DF3 (CV 30) |
| NB | 0.76 | 0.80 | 0.64 |
| DT | 0.58 | 0.59 | 0.60 |
| SVM (linear) | 0.770 | 0.773 | 0.625 |
| SVM (RBF) | 0.533 | 0.531 | 0.60 |
| SVM (POLY) | 0.533 | 0.531 | 0.545 |