Writeup: ANLY 501 Project Part 3

Lujia Deng, Luwei Lei, Janet Liu, Yunfei Zhang

ld781,ll1038,yl879,yz678@georgetown.edu

Monday, December 9, 2019

1 Sentiment Analysis

In the *Broadway Reviews* data set, there are 4,649 reviews written by critics. Having looked through the reviews, we observed that some of the reviewers did not convey their opinions in a straightforward way and that there were multiple sentiments involved in one single review, such as the one in (1) below.[1]

(1) "***The production certainly fulfills its modest creative aspirations***. ***The actors are very good at being bad*** and are so daring with the outrageous physical comedy that we often fear for their safety." – Frank Scheck, Hollywood Reporter, 4/2/2017

To examine whether models could identify the sentiments involved in these reviews, we applied sentiment analysis and converted the rating score as labels. However, it is worth noting that the ratings might not precisely reflect the sentiment. Though it might make sense that the review in (1) was a zero rating, the review in (2) was a 8-score review, as shown below:

(2) "Harry Potter Dances Up Corporate Ladder in 'Business', "As for the appealing Radcliffe, he's eager to please but lacks a certain urgency that makes Finch dangerous and irresistible at the same time. ***He's no singer*** (""I Believe in You,"" the show's best-known song, barely makes an impression) ***and not much of a dancer***. Still, he does both more than respectably in the rousing ""Brotherhood of Man"" finale, which sends us home in a forgiving mood." - Jeremy Gerard, Bloomberg News, 3/27/2011

To get a better sense of what is going on with these reviews and their sentiments, three of us manually annotate 698 reviews (15% of the data set) from our *Broadway Review* data set, and compared models' predictions to both the rating scores and the annotated labels in the test data set. As for label conversion, we converted ratings 0–6 into the NEGATIVE category, 7–8 into the NEUTRAL, and 9–10 into the POSITIVE. Based on our three annotators' annotation experience, we came to the conclusion that reviews labeled as 7–8 ratings in this data set are generally vague, and therefore we set 8 and 6 as thresholds.

[1] The highlighted pieces are what is relevant to the discussion here.

With regard to the preprocessing step, we tokenized the texts, removed stop words, and lemmatized and vectorized the tokens as input. For this sentiment analysis, both logistic regression and Support Vector Machine (SVM) classifiers were employed. Generally speaking, we conducted two sets of classification tasks: *multivariate classification* and *binary classification.* The rationale behind the *multivariate classification* is intuitive: we would like to see how good/bad each classifier performed on the data set. As pointed out above, the sentiments involved in the reviews were not always transparent and easy to capture, and thus we did not expect the model to perform very well. In addition, as demonstrated by the examples in (1) and (2), the NEURAL category could be biased and thus poses noises to the whole classification, we therefore also conducted a *binary classification*; that is, we dropped the data with NEUTRAL labels and let the models perform a binary classification task to compare the results.

1.1 Results

Here we report four sets of results that have demonstrated the best performance among different data and model setups. Each table is provided in the caption a brief description of the input data as well as the model being used.

Multivariate Classification In this classification task, both the logistic regression and SVM models were used to predict on the three categories we identified from the data set, namely, NEGATIVE, NEUTRAL, POSITIVE. Table 1 provides an overall accuracy result on each classifier. Specifically, one can see that the logistic regression classifier outperformed SVM on labels generated from ratings whereas SVM outperformed logistic regression on manually annotated labels. Table 2 and Table 3 show precision, recall, and F1 scores for each scenario. It is interesting to see that the category NEUTRAL was better predicted in the logistic regression model using labels generated from the ratings (see Table 2) and that the category NEGATIVE was better predicted in the SVM model using manually annotated labels (see Table 3).

The models did not have very ideal performances, with 0.53 accuracy in terms of labels from rating scores and 0.41 in terms of manually annotated labels. Hence, we tried to simplified the task by removing the NEUTRAL labels, leading to a binary classification task.

	Logistic Regression	SVM
Accuracy on Ratings	**0.53**	0.49
Accuracy on Annotated Labels	0.38	**0.41**
Avg. Time Cost (min)	0.46	8.65

Table 1: Accuracy on Multivariate Classification.

	Precision	Recall	f1-score
Negative	0.51	0.40	0.45
Neutral	0.56	0.71	**0.62**
Positive	0.46	0.30	0.36
accuracy			0.53
macro avg	0.51	0.47	0.48
weighted avg	0.52	0.53	0.52

Table 2: Logistic Regression Performance on Labels generated from Ratings (Multivariate).

	Precision	Recall	f1-score
Negative	0.47	0.46	**0.46**
Neutral	0.28	0.59	0.38
Positive	0.69	0.29	0.41
accuracy			0.41
macro avg	0.48	0.45	0.42
weighted avg	0.53	0.41	0.41

Table 3: SVM Performance on Manually Annotated Labels (Multivariate).

Binary Classification In this classification task, we dropped the data with Neutral labels and asked the models to perform a binary classification task to make predictions on the Negative and Positive labels. Table 4 provides an overall accuracy result on each classifier. Similar to what have been seen in the multivariate classification task, the logistic regression classifier performed better on labels generated from ratings than SVM whereas worse on on manually annotated labels than SVM, though the difference in each scenario is not big. Table 5 and Table 6 show precision, recall, and F1 scores for each scenario. It is interesting to see that the category Negative was better predicted in the logistic regression model using labels generated from the ratings (see Table 5) and that the category Positive was better predicted in the SVM model using manually annotated labels (see Table 6).

	Logistic Regression	SVM
Accuracy on Ratings	**0.76**	0.73
Accuracy on Annotated Labels	0.66	**0.67**
Avg. Time Cost (min)	0.09	1.86

Table 4: Accuracy on Binary Classification Task.

	Precision	Recall	f1-score
Negative	0.80	0.78	**0.79**
Positive	0.70	0.73	0.72
accuracy			0.76
macro avg	0.75	0.75	0.75
weighted avg	0.76	0.76	0.76

Table 5: Logistic Regression Performance on Labels generated from Ratings (Binary).

	Precision	Recall	f1-score
Negative	0.53	0.77	0.63
Positive	0.82	0.61	**0.70**
accuracy			0.67
macro avg	0.68	0.69	0.66
weighted avg	0.71	0.67	0.67

Table 6: SVM Performance on Manually Annotated Labels (Binary).

1.2 Discussions

As can be seen from 1.1, models produced better performances in a binary classification task, with logistic regression model performing better than SVM on labels generated from ratings. The results indicate that the model was able to predict extreme sides of the review to some extent whereas reviews that were vague in meanings were difficult for models to learn.

In addition, we also observed several issues that might have led to inaccurate predictions:

- High ratings do not always correspond well with the reviews (e.g. the rating of 8 is not consistent across the board);
- Since these labels were annotated by three annotators, there will be biases across the board. That is, different annotators will have different criteria.
- Some reviews did not provide a strong sentiment towards the shows; but instead, these reviews were centered on the critic of a role, which made it very difficult to decide its sentiment.

2 Topic Modelling

In addition to conducting sentiment analysis on critics' reviews, we also performed a topic modeling analysis to capture the topics that are prevalent among the critics. We treated each review as one document and applied latent Dirichlet allocation (LDA) model to the overall texts. Having experimented the model and observed the performance with different numbers of topics, we decided to set the number of topics to four. Figure 1 demonstrates the distribution of document word counts by topics. Figure 2 shows the top ten keywords in each topic. The size of the words are proportional to their weights in the word clouds.

As can be seen from Figure 1, Topic 2 has the highest number of documents, but it also has more general keywords such as "good", "people", and "make". Topic 1 and Topic 3 provide more meaningful insights even though they have lower numbers of documents. Moreover, as can be seen from Figure 2, critics were focused on the themes of the shows such as "friend", "death", and "woman"; also, we are able to identify some synonyms such as "comedy", "laugh", and "funny".

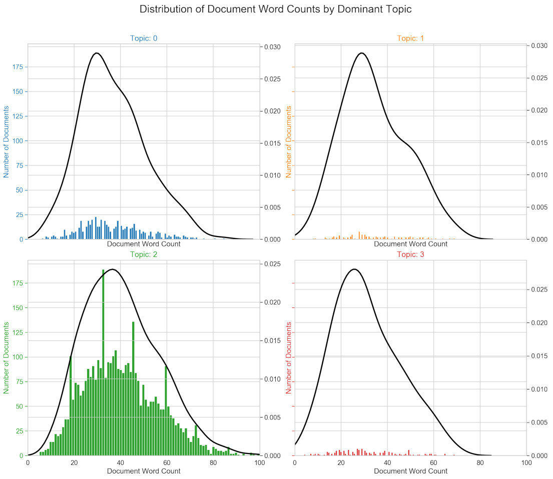

Figure 1: Document Distributions for Each Topic.

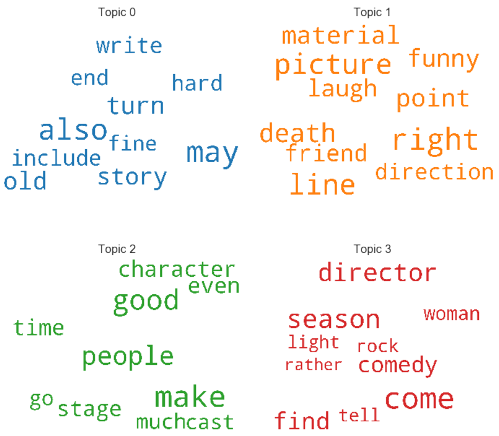

Figure 2: Topic Keywords in Each Topic.

3 Limitations

There are some limitations we have identified regarding our data collection, data processing, and analysis.

3.1 Limitations of Data

Firstly, our data sources are not very diverse and transparent. All of our data sets are from Playbill.com and broadwayworld.com. It would be better if we would have considered other data soruces. Secondly, when collecting the data, we accidentally excluded theater information. As a result, it is likely that we have lost some important and valuable information that could impact our analyses. Thirdly, when aggregating our data with regard to the dates to represent seasonality, we came up with new attributes such as "Thanksgiving Week", "Winter Holiday Break", and "Summer Break". It is likely that the cutoff points of each period is not highly accurate across the board.

3.2 Limitations of Analysis

In addition to the limitations that we have discussed in 1.2 with regard to sentiment analysis, we also identified the following limitations of our analysis.

Firstly, we did not use the *Broadway News* data set for any type of analysis sine all our textual analysis has been done on the *Broadway Review* data set. It would be interesting to see what topics or sentiments these news have and how these news information could be collated with other data sets. Secondly, although we have tried to established solid connections between shows' reputation/popularity and the factors associated with them, it is still not very clear-cut to derive such generalizations due to the nature/quality of the data we have.

4 Ethical Considerations

Although there are no salient ethical issues associated with our data collection and analysis, we were inspired by the discussion from the last class. Therefore, we point out some of the ethical considerations we come up with for the project.

4.1 Data Sources

Our data is from Playbill.com and broadwayworld.com, which are two platforms directly or indirectly affiliated with Broadway. As a result, these data could be biased, and we will be criticized in terms of not using more transparent data due to profit interests. To better support or protect this interest at risk, we could diversify our data sources.

4.2 Impact on Broadway Theaters & Producers

The analyses we have provided could potentially impact Broadway Theaters' decisions on scheduling shows based on their reputation or popularity. In addition, our analyses could also provide insights into how producers design / structure their future shows.

4.3 Impact on Tourists & Ticket Buyers

When preprocessing our data, we eliminated shows with preview but without performance information. This could potentially exclude shows that could achieve a high popularity in the future. As a result, tourists or ticket buyers who try to gauge some expectations based on our analyses could be affected as they are not guaranteed to have a whole picture.

5 Findings & Conclusion

To conclude, having conducted several exploratory and predictive analyses in this project, we identified some factors that may or may not be indicative of a shows popularity and grosses such as number of performance in a week, social media reactions, theater size, and average ticket price etc. We also identified the impact that seasonality has had on Broadway shows. This important factor cannot be directly observed from the data sets we collected. However, with our hypothesis testing as well as the classification task, we have obtained some insights that explain our intuition. In addition, when we observed some abnormal fluctuations in the data sets as well as our analysis, we did some research to explore the historic background underlying the phenomena.

Additionally, we also found out that ***Thanksgiving*** is a more profitable season for Broadway comparing to the rest of the year while summer time is less significant. Moreover, ***theater size*** has a positive relationship with the ***average ticket price***. And the rating of a show doesn't contribute to the weekly gross. The predictive model also helps to suggest that the most important social statistic features are ***Date*** and ***Facebook Checkins***.

Considering the context surrounding the analysis, we need to say that although we are trying very hard to establish solid/legitimate connections between a shows reputation or popularity and its associated factors from different data sets we have, it is still not very clear to draw such conclusions. We certainly do not want to make sweeping generalizations, but our analysis should lay fundamental foundations that can spark further research along this line of interdisciplinary study.