Generalizing Sentiment Analysis Techniques Across Sub-Categories of IMDB Movie Reviews

Nick Hathaway

Advisor: Bob Frank

Submitted to the faculty of the Department of Linguistics in partial fulfillment of the requirements for the degree of Bachelor of Arts

Yale University

May 2018

TABLE OF CONTENTS

Abstract

Natural language processing systems have had to deal with differences across linguistic genres. Systems trained on data from one domain do not necessarily perform well on data from another domain. Yet, it is not completely clear how domains should be defined for any given task. In this senior essay, I investigate this issue in the context of sentiment analysis, a task which identifies the polarity of a given text (for example, whether a movie review is positive or negative).

I explore the usefulness of dividing a corpus of movie reviews drawn from the Internet Movie Database (IMDb) into five different genres: animation, comedy, documentary, horror, and romance. I demonstrate that sentiment generalizes well across genres with the exception of horror movie reviews, which tend to contain a higher proportion of negative words in both positive and negative reviews. As a result, training on horror movies will lead to a sentiment analysis system with higher precision but lower recall in other domains.

Acknowledgements

First, many thanks to my advisor, Bob Frank, for his guidance and support throughout my writing process. I could not have written this paper were it not for our meetings.

Thank you to my professors in Yale's Linguistics Department for their fascinating courses, helpful assignments, and challenging final projects. I also want to thank the graduate students in the department who have given me advice and have been useful resources during the past couple years.

And thank you to Raffaella Zanuttini as well as my fellow linguistics seniors for your moral support and feedback.

1. Introduction

In natural language processing systems, models trained on one domain do not necessarily generalize well to data from other domains. It is often unclear how much a domain should be restricted in order to create robust models. In the domain of tweets, examples are often drawn from a stream of many tweets from many sources. In movie reviews, natural language corpora are defined to include all movies, regardless of their possible sub-categories. Underlying these decisions is the assumption that sentiment can be effectively abstracted without restricting the domain to more specific subdomains (for example, to movie genres or by release date).

In order to investigate this claim, I will build a corpus of movie reviews divided into five different genres: animation, comedy, documentary, horror, and romance. I will then train an SVM binary classifier on training and testing splits that will demonstrate the generalizability of the model with respect to movie genre. If the genres have similar outcomes to their peers, that will provide evidence that the features predicting sentiment in a movie review can be extracted from any type of movie, regardless of genre. If a classifier trained on one of the genres performs poorly on the others, then there might be reason to believe that sentiment can have genre-specific qualities.

User-generated movie reviews found on sites like The Internet Movie Database (IMDb) and Rotten Tomatoes have become popular sources of data to investigate questions in many subfields of computational linguistics. The sentiment analysis models trained on this data have a wide range of applications. Web scraping has made it possible to generate large corpora of natural language data across many domains. However, all of these efforts lead to a more complex question: How do we define a domain? Or more specifically, to build a sentiment analysis corpus, should we collect all types of reviews, movie reviews, horror movie reviews, or B-movie horror movie reviews? Through the lens of sentiment analysis in the domain of movie reviews, I will examine the generalizability of classification models across movie genres.

I begin with an overview of other sentiment analysis corpora that have online movie reviews, specifically Pang and Lee's Internet Movie Database (IMDb) corpus (Pang and Lee 2004) and Maas et. al's IMDb corpus (Maas et. al 2011). Then, I discuss my own approach to building a corpus by scraping and cleaning user-generated movie reviews from IMDb to create a corpus of approximately 1.2 million usable reviews. By analyzing the corpus, I identify potentially useful features and investigate potential difficulties with the data, such reviews for the same movie titles. In my methodology and results sections, I summarize the results of the four different training-testing splits, focusing mainly on how the horror genre differs from its peer genres. Finally, I discuss the pitfalls of using movie genre as a subdomain and further research questions stemming from my conclusions.

2. Overview of Movie Review Datasets

To test my hypothesis, I will build a movie review corpus subdivided by genre by scraping The Internet Movie Database. Then, I will examine trends across the five genres in my corpus to identify potential properties that might make some genres more robust than others.

Before building the corpus, I looked at review datasets used in sentiment analysis to identify their sources, length, format, rating cutoffs, among other features. Specifically, I looked at Pang and Lee's IMDb corpus (2002) and Maas et. al's IMDb corpus (2011). While I mainly discuss these two datasets, I also looked at Nguyen etl. Al's IMDb corpus (2014) and Blitzer et. al's Amazon review corpus (2007) for guidance when building my own review corpus.

Pang and Lee's Cornell Polarity Data v2.0 (Pang and Lee 2004) movie review corpus contains 2,000 labelled IMDb reviews (1,000 positive and 1,000 negative). All reviews were written before 2002 and no more than 20 reviews were scraped per author. Overall, the corpus has reviews written by 312 different authors. The reviews are split into different sentences on each line and all references to the rating of the review have been removed. They have filenames that indicate how they were scraped from the accompanying html. To determine the positivity or negativity of each review, Bo Pang and Lillian Lee only considered ratings with a maximum (i.e., 4 out of 5 stars, 8/10). They defined the following positive and negative thresholds:

- Five-star system:
    - Positive: >= 3.5 stars
    - Negative: <= 2 stars
- Four-star system:
    - Positive: >= 3 stars
    - Negative: <= 1.5 stars
- Letter grade:
    - Positive: >= B
    - Negative: <= C-

Maas et. al's (Maas et. al 2011) corpus contains 50,000 labelled IMDb reviews (25,000 positive, 25,000 negative) and an additional 50,000 unlabelled IMDb reviews. They allowed no more than 30 reviews per movie to avoid correlated ratings. Each review's filename contains that file's unique id and its rating. They have included a list of all imdb review pages for each of the movies used to build out the corpus. In addition to the plain review data, they included a tokenized bag of words list of features. They defined the positivity/negativity of reviews differently for their labelled and unlabelled sets:

- Labelled (ten-star system):
    - Positive: >= 7 stars

    - Negative: <= 4 stars
- Unlabelled (ten-star system):
    - Positive: > 5 stars
    - Negative: <= 5 stars

While these two corpora have many differences, especially with regards to their size, there are two details that I believe to be the most important: Maas et. al's corpus allowed no more than 30 reviews per movie and both used the same positive and negative rating thresholds to define their polar categories (4 or less for negative reviews, 7 or more for positive reviews on a 10 point scale). I used the same thresholds in my corpus. However, I chose to not limit reviews by movie title. By doing this, I was able to also look at the expected number of reviews per movie pulled from a random sampling of reviews.

3. Building the Corpus

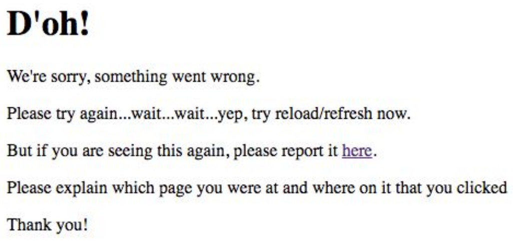
In order to evaluate the generalizability of sentiment analysis techniques over subdomains of movie reviews, I used a combination of movie review APIs and web scraping libraries to build a corpus of around 130,000 movie reviews. The APIs that I used support generating lists of movie titles across many different filters, including MPAA film rating, box office, release date, and genre. The first corpus I generated was subdivided by movie genre. However, my methods can easily extend to any of the above criteria (or to a custom list of movies).

To generate movie titles and divide them into different genres, I used an API from The Movie Database's (TMDb).[1] I searched for movies from five different genres: Animation, Comedy, Documentary, Horror, and Romance. Unfortunately, TMDb limits their search features to the first 20,000 results, which put an artificial cap on the number of movie titles in the corpus.

After generating lists of movie titles, I used the Open Movie Database (OMDb) API[2] to translate them into unique IMDb IDs. This stage of the corpus generation process caused some genres to lose as much as 40% of their overall size (14,357 animated movie titles resulted in only 8,971 unique IMDb IDs).

[1] This resource can be accessed at https://www.themoviedb.org and used with an API key.

[2] The second API I accessed can be found at http://www.omdbapi.com/



D'oh!

We're sorry, something went wrong.

Please try again...wait...wait...yep, try reload/refresh now.

But if you are seeing this again, please report it here.

Please explain which page you were at and where on it that you clicked

Thank you!

Using the five lists of IMDb IDs, I scraped all of the reviews for each movie from IMDb. Instead of generating each page of reviews using user-facing urls,[3] I decided to scrape the movie review dating using their _ajax urls.[4] This approach is more robust to site redesigns, which are often used to prevent web scrapers (IMDb recently redesigned their website in November, 2017). Each _ajax url has an optional pagination key allowing the script to scrape all of the reviews for a movie in order.

In order to avoid being blacklisted by IMDb for sending too many requests at a time, I scheduled each request to send every 0.5 seconds. For each movie, I created a .txt file with that movie's IMDb ID as the title. These files contain limited metadata about the movie itself (name, IMDb ID, and total number of reviews), but the scraping script can be easily modified to include additional information about each movie title. Each review has a title, publication date and rating, followed by the full text of the review.

[3] User friendly urls are the actual pages you might see if you were to search for reviews of a movie on IMDb (for example, http://www.imdb.com/title/tt0077766/reviews)

[4] IMDb's _ajax urls produce the review content without any styling or redundant information (http://www.imdb.com/title/tt0077766/reviews/_ajax), making web scraping more robust and efficient

Movies that returned zero reviews were removed from the corpus, as well as non-English movies. Additionally, reviews with N/A ratings were also removed by the corpus. Cleaning the initial IMDb reviews brought the size of the corpus from 61,265 titles to 42,037 titles. Overall, this process resulted in a corpus of reviews (positive, negative, and unrated) for 42,037 movies across the five genres.

Genre	**TMDb Search (titles)**	**OMDb Search (IMDb IDs)**	**Number of Movies Scraped (Overall)**	**Number of Movies (no empty files)**
Animation	14,357	8,971	8,276	5,483
Comedy	20,000	15,500	14,454	11,052
Documentary	20,000	12,862	12,396	7,222
Horror	17,932	14,214	12,109	8,677
Romance	20,000	15,514	14,030	9,603
ALL	92,289	67,061	61,265	42,037

Figure 1: Number of Movies Reduced At Each Stage of Scraping

After generating the corpus, I filtered out reviews with no ratings and divided them into positive and negative categories (7 or more stars being positive, 4 or less being negative). The resulting corpus contained separate files for each movie (just as before, they were named after their unique IMDb ID). Each file contains only the review text.

4. Corpus Analysis

The final version of my corpus shows that, in general, positive reviews are more common than negative reviews. Across genres, there are usually four times as many positive reviews as negative reviews (except in horror, where there are only two times as many positive reviews as negative reviews). Additionally, there are usually a similar number of negative and unrated reviews (except, again, in horror, where there are approximately 50% more negative reviews than unrated reviews).

Genre	Class	Number Movies	Number of Reviews
Animation	Positive	4,578	120,947
	Negative	2,628	34,751
	Unrated	2,977	32,889
Comedy	Positive	9,855	252,001
	Negative	7,440	85,300
	Unrated	7,369	79,160
Documentary	Positive	6,191	118,957
	Negative	3,049	35,475
	Unrated	3,611	32,469
Horror	Positive	7,164	203,467
	Negative	6,441	97,081
	Unrated	5,402	65,541
Romance	Positive	8,394	222,517
	Negative	6,011	68,829
	Unrated	6,124	65,189

Figure 2: Final distribution of reviews

In addition to the distribution of positive and negative reviews by genre, there were patterns in the length of reviews and word length and usage across genres. For example, positive reviews tend to be longer than negative reviews. In general, comedy reviews are shorter than their genre peers. Perhaps positive reviews are longer because reviewers are more willing to put time into reviews of movies they enjoyed. Given this difference in length of review by positive category, it is especially important to adjust positive and negative word counts by overall review length (percent_pos_words and percent_neg_words in my list of features).

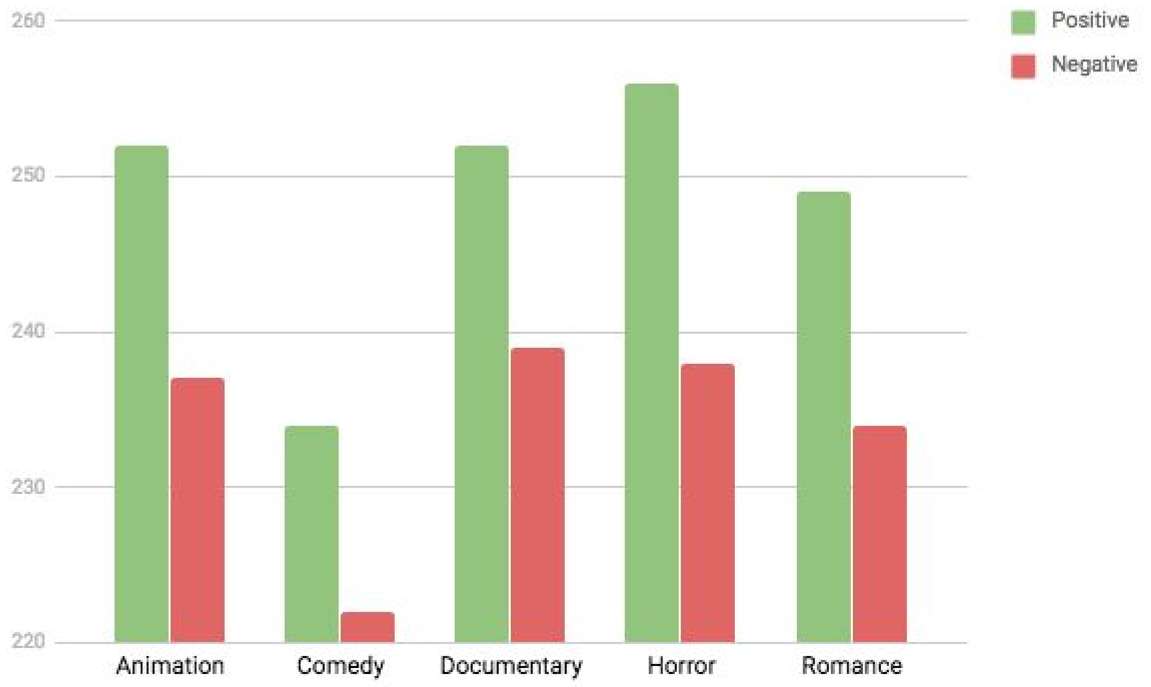


Figure 3: Average length of review by genre and polarity

Another interesting trend is that positive reviews tend to use longer words than their negative counterparts. However, the differences were too slight to be an extracted feature in the SVM model. Note that all of my corpus analysis was completed after stemming and removing the stop words from the reviews. This is because I wanted to determine which features might make informative contributions to the SVM classification model.

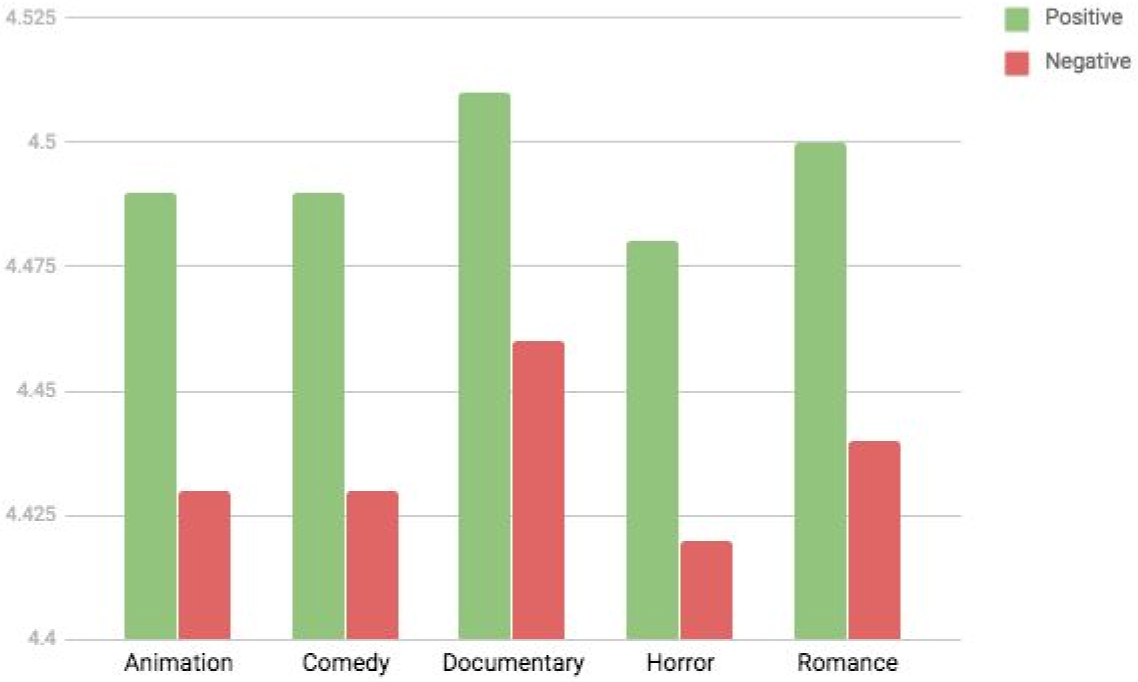


Figure 4: Average word length by genre and polarity

I decided to count the number of positive and negative words per 100 words for each genre (again, on reviews with stop words removed.) It is interesting to note that for positive reviews, there are around twice as many positive words as there are negative words. However, for negative reviews, there are roughly the same number of negative and positive words across all the reviews. This trend is true across all genres as well, with slight variations. For example, horror has fewer positive words and more negative words in both its positive and negative reviews and romance reviews have more positive words and fewer negative words compared to the other genres. These features might be influenced by plot summaries in the reviews in addition to the subjective opinions of the author.

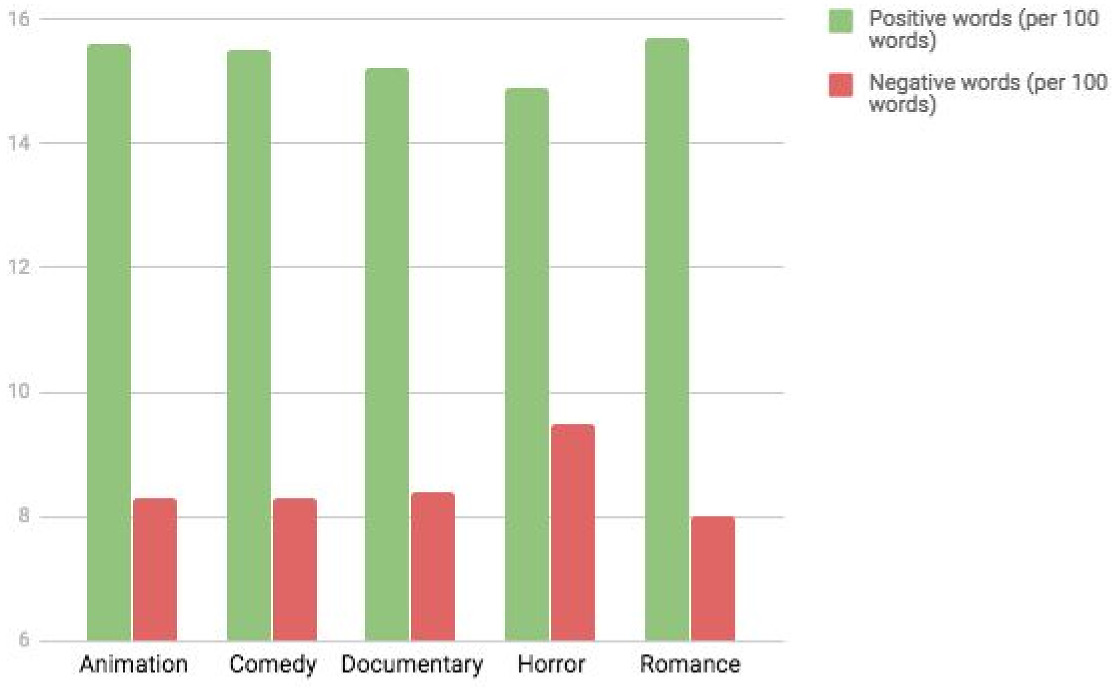


Figure 5: Polar word counts for positive reviews by genre

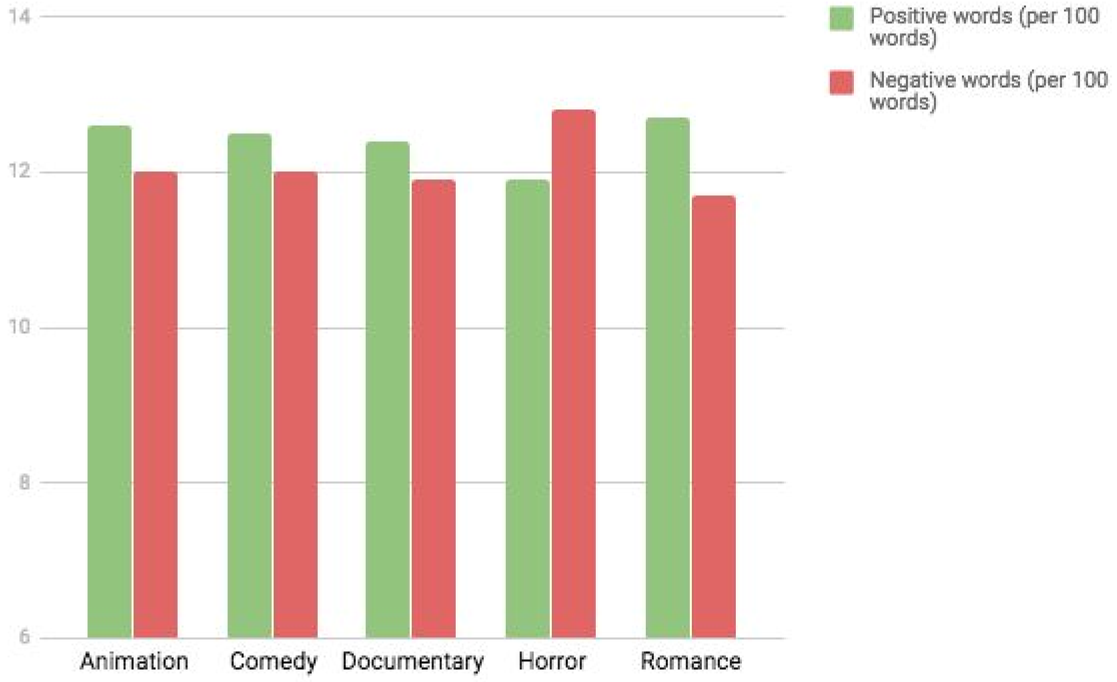


Figure 6: Polar word counts for negative reviews by genre

The titles in my corpus range from contributing one review to contributing hundreds. The distribution of review counts per title is logarithmic, with the majority of movies only contributing 1-5 reviews (see Figure 7 on the next page). The overall trend is even clearer when review counts are not grouped together (as in the 31+ category below). However, I decided to cut off the review counts after 30 because of Maas et. al's (Maas 2011) decision to restrict reviews to only 30 per movie title in their IMDb corpus.

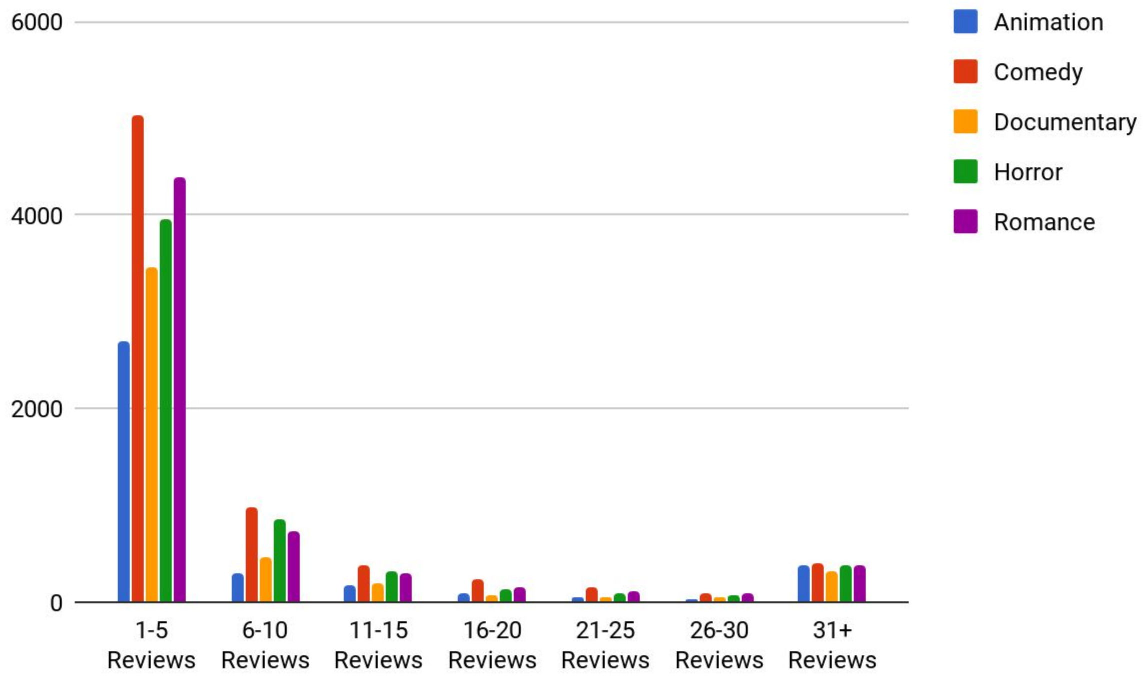


Figure 7: Number of movie titles by overlapping review contributions

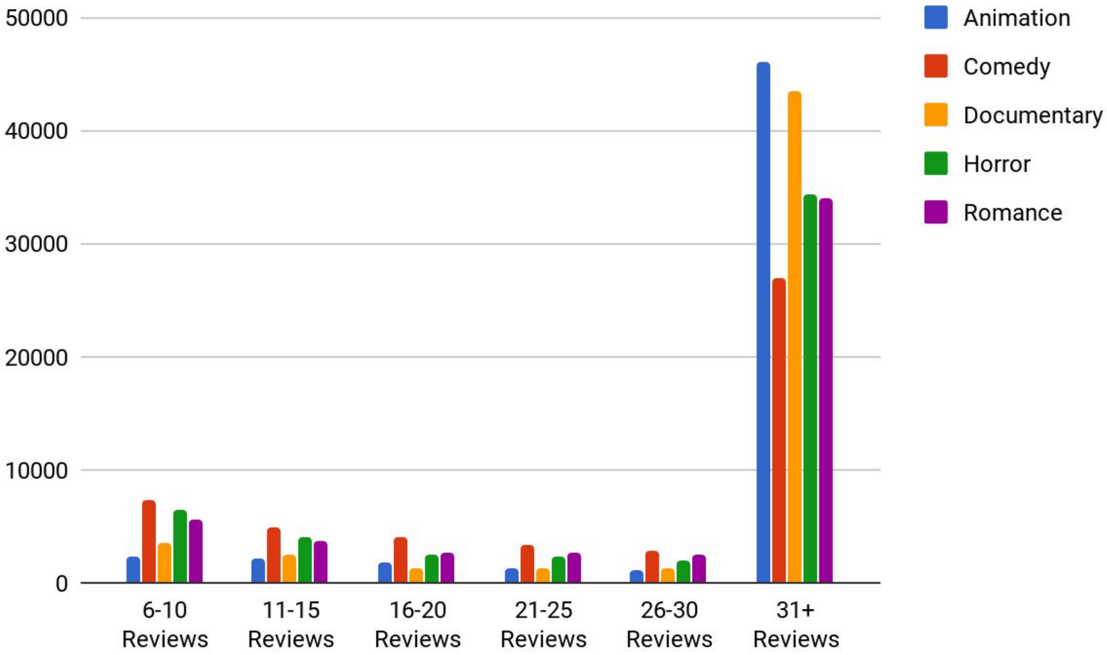


Figure 8: Number of total reviews by review contribution

When these categories are adjusted by the number of reviews they actually contribute, it is clear that the majority of reviews come from movies with more than 30 reviews, as in Figure 8.

When these counts are adjusted to the actual number of reviews contributed per review count category (as in Figure 9), we see that there is a fairly even chance of selecting a review with any number of reviews for the same title. Any randomly selected movie title will come from a movie with approximately 147 total reviews in the corpus on average (across all genres). Additionally, the higher counts for overlapping reviews are much sparser than the lower counts. Some only have one movie title in their category, yet contribute hundreds of reviews to the corpus.

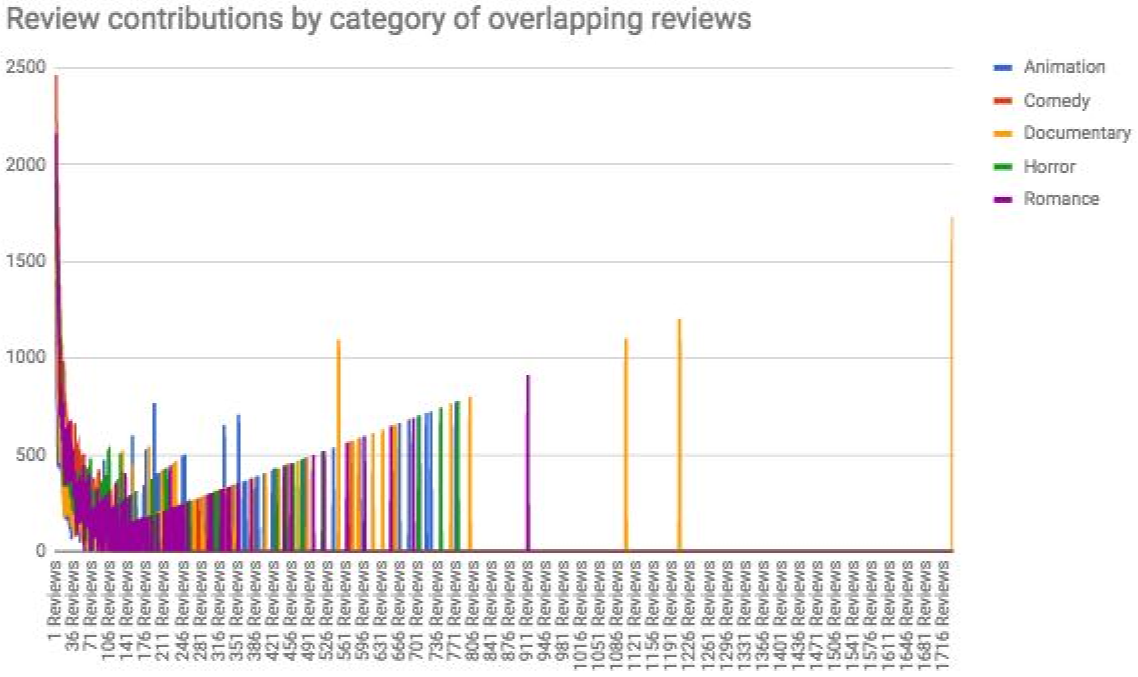


Figure 9: Review contributions by category of overlapping reviews (all categories)

5. Methodology

Because the reviews in each of the genres skew positive, I have decided to limit the final corpus to 60,000 reviews from each genre (30,000 positive and 30,000 negative). In order to randomly select the 300,000 reviews, I splitting each unique movie file into separate files for each review (in the form tt008329-1.txt, tt008329-2.txt, etc.). I have decided to not limit the number of reviews by movie.

First, I trained on an svm classifier on a corpus of mixed genre reviews and tested the classifier on held out data from each of the five movie genres. In order to obtain an ideal training-testing split, I held out 150,000 mixed genre reviews for training and 150,000 reviews for testing (across the five genres). The classifier is trained once on the 150,000 review mixed genre corpus and then tested five different times on each genre's 30,000 review corpus. This results in a 83.3% training, 16.7% testing split.

Genre	Training	Testing
Animation	150,000	30,000
Comedy		30,000
Documentary		30,000
Horror		30,000

Romance		30,000

Figure 10: Training on a mixed category corpus and testing on a specific subgenre

These results should demonstrate the generalizability of models trained in the overall domain of movie reviews. I will compare its performance on each of the five genres to discover if the mixed genre model is robust. Because sentiment analysis techniques often train on movie reviews in general without regard for their genre, differing results across genres would suggest that reducing domains into more specific subcategories could increase performance.

After testing the mixed genre classifier on specific subgenres, I will run the svm classifier in the reverse direction. In other words, I will train five separate svm classifiers using 57,500 review corpora from each genre. Then, I will test each of the classifiers on a mixed genre corpus consisting of 12,500 reviews. This results in a 82.1% training, 17.9% testing split.

Genre	Training	Testing
Animation	57,500	12,500
Comedy	57,500	
Documentary	57,500	
Horror	57,500	

Romance	57,500	

Figure 11: Training on specific subgenres and testing on a mixed category corpus.

Results from this phase of classification will demonstrate which movie genres create the most generalizable model when tested on a mixed genre corpus. If any of the genres outperform the others, future models might take advantage of their superior generalizability. For example, if the classifier trained on the comedy genre outperforms the other genres, we could rely more on comedy movie reviews in future applications, especially if they are more abundant.

In addition to these two training/testing splits, I will also train classifiers on each individual genre and test them on the other four genres. This will better demonstrate the differences across genres by showing how well a model trained on one genre can predict the positivity or negativity in reviews from other genres. Lastly, I will look at a classifier trained on four genres and tested on one genre. Doing this will prevent the overlap found by training on a mixed dataset by removing any reviews from the genre that will be used to test the classifier.

To test their generalizability, I used a Linear SVC classification model. I extracted the following feature sets from each review:

- Unigram word counts

- Polarity word counts (adjusted to review length)
- Ratio of positive to negative and negative to positive words
- Bigram counts

I determined which unigram words to include by identifying the top 500 most frequent words of all the review texts once they had been stemmed and stop words had been removed. For bigrams, I looked at the 100 with the highest mutual information that had occured at least three times in the corpus. Positive and negative words were pulled from a stemmed version of Hu and Liu's sentiment lexicon (Hu and Liu 2004). Their lexicon contains around 6,800 positive and negative words, evenly split. However, this count includes all part of speech variants of the same stemmed word (i.e., "disgrace," "disgraced," "disgraceful," and "disgracefully.")

6. Results

To evaluate my models, I have decided to use a balanced F-1 Score. This ensures that the evaluation of each classifier's performance depends on a joint measure of precision and recall. Neither precision nor recall are weighted higher than the other.

$$2\ \frac{Precision * Recall}{Precision + Recall}$$

From the below list of most informative features, we can see that the negative word counts have little impact on predicting a review as negative. By contrast, positive word counts have a large effect on labelling a review as positive (a coefficient of 0.6991). This follows from the charts from the Corpus Analysis section of my paper, which showed that negative reviews have the same number of negative and positive words on average, whereas positive reviews have around twice as many positive words as negative words. Because the feature extraction only considers word counts and not their contexts, it is likely that plot summaries have contributed negative words to positive reviews.

Negative		**Positive**	
Word	**Coefficient**	**Word**	**Coefficient**
Worst	-0.4315	Percent_pos_words	0.6991
Wast	-0.4329	Highli recommend	0.5533
Aw	-0.3886	Must see	0.4285
Bore	-0.2835	Edge seat	0.4267
Terribl	-0.2659	Top notch	0.3735
Unfortun	-0.2602	Excel	0.2827
Suppos	-0.2370	One best	0.2750
Horribl	-0.2163	Perfect	0.2694
Disappoint	-0.2139	Brilliant	0.2324
Poor	-0.2107	Hilari	0.2280

Wors	-0.2095	Favorit	0.2061
Fail	-0.2018	Laugh loud	0.2026
Noth	-0.1975	Amaz	0.1750
Predict	-0.1948	Definit	0.1567
Ridicul	-0.1856	Fantast	0.1566
Save	-0.1776	Well done	0.1562
Wast time	-0.1753	Touch	0.1398
Instead	-0.1692	Even though	0.1386
Attempt	-0.1685	Ever made	0.1359
Lack	-0.1641	Worth watch	0.1355

Figure 12: Top 20 most informative features predicting positive or negative reviews

Looking at the top 20 most informative features for the mixed genre classifier, there are a couple clear trends. First, the features that predict negativity are almost always single words and not bigrams, with the exception of “waste time.” There are no non-unigram or non-bigram features in the list predicting negativity. As for the features predicting positivity, the most informative feature was percent_pos_words, or the count of positive words divided by overall review length. About half of these features are bigrams. For example: highly recommend, must see, edge seat, top notch, one best, and laugh loud. It’s important to note that some of these bigrams would not have been captured if stop words were not removed first.

The most informative features for classifiers trained on a specific genre show the same general pattern. Animation included both percent_pos_words and percent_neg_words while documentary included percent_pos_words in their top 20 features for predicting positive reviews. The only genre specific bigram predicting negativity was “soap opera” (-0.2066 coefficient) for romance reviews. The other bigrams were usually movie titles and actors names, such as “looney toon” (0.2432) for animation and “samuel l” (0.1895) and “elm street” (0.1613) for horror.

Despite looking at polar word counts and ratios, they did not rank in the top 20 most informative features for any of the classifiers. Even in the mixed dataset, where the percent_pos_words feature had a positive coefficient of 0.6991, even one occurence of the bigram “highly recommend” provides almost as much predictive power with its 0.5533 coefficient.

Overall, the classifiers performed similarly across genres. Almost all of the different training and testing splits had F-Scores ranging from around 84% to 86%. While there were slight differences in precision and recall across genres, there are no trends significant enough to show that some genres generalize better than others.

Genre	**Polarity**	**Precision**	**Recall**	**F-Score**
Animation	Positive	86.77	84.49	**85.61**
	Negative	84.88	87.11	**85.98**
Comedy	Positive	86.61	83.48	**85.02**
	Negative	84.06	87.09	**85.55**
Documentary	Positive	85.96	83.17	**84.55**
	Negative	83.70	86.42	**85.04**
	Positive	88.26	81.05	**84.50**

Horror	Negative	82.48	89.22	**85.72**
Romance	Positive	86.04	85.23	**85.64**
	Negative	85.37	86.17	**85.77**

Figure 13: Trained on mixed genres, tested on specific genres

However, there were consistent differences in precision and recall for horror reviews. When trained on the mixed dataset and tested on horror, the classifier had a higher precision (88.26%) and lower recall (81.05%) for positive reviews and a lower precision (82.48%) and higher recall (89.22%) for negative reviews. This shows that the mixed genre classifier often incorrectly labeled positive horror reviews as negative, resulting in the higher precision and lower recall for positive reviews. Positive horror reviews are more likely to contain a higher number of negative words compared to positive reviews from other genres. This may be due to the inclusion of plot summaries in all of the movie reviews, with horror movies' plots being more likely to contain negative events.

Genre	**Polarity**	**Precision**	**Recall**	**F-Score**
Animation	Positive	84.92	86.02	**85.46**
	Negative	85.83	84.72	**85.27**
Comedy	Positive	84.69	86.74	**85.70**
	Negative	86.41	84.32	**85.35**
Documentary	Positive	85.72	84.13	**84.92**
	Negative	84.42	85.98	**85.19**
Horror	Positive	82.04	88.48	**85.14**
	Negative	87.50	80.62	**83.92**
Romance	Positive	85.44	85.15	**85.30**
	Negative	85.20	85.49	**85.34**

Figure 14: Trained on specific genres, tested on mixed genres

When the classifier was trained on horror reviews and tested on the mixed dataset, it resulted in the opposite: lower precision and higher recall for positive reviews and higher precision and lower recall for negative reviews. Specifically, the classifier had a 82.04% precision and 88.48% recall for positive reviews and a 87.50% precision and 80.62% recall for negative reviews. For both of these training and testing splits, there was around a 6% to 7% difference in their precision and recall. The classifier trained on horror reviews required a higher number of negative words in a review for it to be considered negative (relative to negative reviews from other genres). This caused a higher precision and lower recall for negative reviews because the reviews classified as negative met the classifier's higher threshold for negative classification. Similarly, the classifier also identified negative reviews as positive because positive horror reviews also contain more negative words compared to positive reviews from other genres.

The classifier trained on the four other genres and tested on horror showed similar results to the classifier trained on the mixed genre dataset (which included reviews from horror). However, the differences in precision and recall are slightly less pronounced (ranging from 4% to 5%).

Genre	**Polarity**	**Precision**	**Recall**	**F-Score**
Animation	Positive	84.90	87.12	**86.00**
	Negative	86.78	84.51	**85.63**
Comedy	Positive	85.66	85.55	**85.61**
	Negative	85.57	85.68	**85.62**
Documentary	Positive	84.56	85.14	**84.85**
	Negative	85.04	84.45	**84.74**
Horror	Positive	87.311	83.11	**85.24**
	Negative	83.91	88.10	**85.96**
Romance	Positive	83.76	88.73	**86.18**
	Negative	88.02	82.80	**85.33**

Figure 15: Trained on four other genres, tested on specific genres

This pattern for horror reviews is even more evident when the classifier is trained and tested on individual genres. After being trained on horror and tested on the other four genres, the difference between precision and recall ranges from around 6% to 10%. When tested on romance reviews, the classifier had a 81.67% precision and 90.32% recall for positive reviews and a 89.17% precision and 79.72% recall for negative reviews. These results support the same trend shown by the classifier trained on horror and tested on the mixed genre dataset.

Genre	**Polarity**	**Precision**	**Recall**	**F-Score**
Animation	Positive	82.64	88.91	**85.66**
	Negative	88.00	81.32	**84.53**
Comedy	Positive	82.79	88.38	**85.50**
	Negative	87.54	81.63	**84.48**
Documentary	Positive	81.62	87.80	**84.60**
	Negative	86.80	80.23	**83.39**
	Positive	81.67	90.32	**85.78**

Romance	Negative	89.17	79.72	**84.18**

Figure 16: Trained on horror, tested on all other genres individually

When trained another genre and tested on horror reviews, we see the same trend shown by the classifier trained on mixed genres and tested on horror. Specifically, the classifier trained on romance and tested on horror had the most dramatic results: a 89.34% precision and 78.55% recall for positive reviews and a 80.86% precision and 90.63% recall for negative reviews.

Genre	**Polarity**	**Precision**	**Recall**	**F-Score**
Animation	Positive	87.78	81.54	**84.54**
	Negative	82.76	88.65	**85.60**
Comedy	Positive	87.51	82.19	**84.77**
	Negative	83.21	88.27	**85.66**
Documentary	Positive	88.22	79.89	**83.85**
	Negative	81.63	89.33	**85.31**

Romance	Positive	89.34	78.55	**83.60**
	Negative	80.86	90.63	**85.47**

Figure 17: Trained on individual genres and tested on horror

It’s important to note that even though horror reviews show differing precision and recall for negative and positive reviews, they have similar F-Scores compared to the other genres. Looking at the following table for classifiers trained on the horizontal genres and tested on the vertical genres, it’s clear that the joint F-Scores are very similar (between 84.00% and 85.46%).

	Animation	Comedy	Documentary	Horror	Romance
Animation		85.46	85.29	85.10	85.22
Comedy	85.17		85.09	85.00	85.10
Documentary	84.53	84.50		84.00	84.41
Horror	85.07	85.22	84.58		84.54
Romance	85.48	85.52	85.23	85.00	

Figure 18: Trained on horizontal, tested on vertical genres (combined negative and positive F-Scores)

7. Discussion

The results from my analysis suggest that subdividing movie reviews by genre has little to no effect on the performance of a Linear SVC sentiment classifier. Overall, the classifier did not capture any genre-specific features that would have resulted in differing accuracy measures. Despite looking at four training-testing splits, the only genre that showed consistent differences from the other genres was horror. Even then, the classifiers trained or tested on horror still had similar F-Scores compared to the other genres.

Because horror movies contain more negative words on average in both positive and negative reviews, they resulted in larger changes in precision and recall. Classifiers trained on horror reviews had higher precision and lower recall for negative reviews and lower precision and higher recall for positive reviews. The opposite was true when classifiers were trained on other genres or mixed datasets of genres and tested on horror.

The main difficulty in subdividing a movie review corpus by genre is that many review sites will label a movie with multiple genre tags. IMDb, among other popular user-generated review sites, has a wiki model that allows any user to add or edit genre tags. When genres are labeled and corrected according to community input, the result is a large overlap of unrelated movies in any given genre. Coco, a popular 2017 CGI movie created by Pixar, has been tagged on IMDb as being in the following genres: Animation, Adventure, Comedy, Family, Fantasy, Music, and

Mystery.[5] Because genres are amorphous on sites such as IMDb, it is difficult to create a corpus that distinguishes them based on their features.

Taking a closer look at the IMDb genre guidelines, it's clear that some genre divisions are more strictly reinforced than others. The guidelines only describe animation and documentary as "objective," while the other three genres in my corpus are seen as "subjective." Animation is more "objective" because IMDb requires that "75% of the title's running time should have scenes that are wholly, or part-animated" (see "Genre Definitions" under references). For documentaries, the titles must "contain numerous consecutive scenes of real personages and not characters portrayed by actors." Perhaps surprisingly, IMDb recommends that "stand-up comedy and concert performances" also be labeled as documentaries ("Genre Definitions.")

When movies are tagged with multiple genres, it becomes more difficult to divide review domains with distinct boundaries. Overlapping genres for movie titles introduces a new question: are genres additive or do combinations of genres have their own unique fingerprint? For example, when considering a concert recording that has been tagged as both "music" and "documentary", does it belong to a specific category of "music documentaries" or can the positive and negative sentiment in its reviews be adequately described by the "music" and "documentary" genres jointly. Because genre tags are not given individual weights or rankings on IMDb, the genres that are only partially descriptive are given the same importance

[5] This review can be found at https://www.imdb.com/title/tt2380307/

as genres central to the movie’s description. For future investigations, it might be interesting to explore the differences across multiple genre categories.

8. Conclusion

Sentiment generalizes well across all genres except for horror movie reviews due to their higher percentage of negative words in both positive and negative reviews. This shows that the correct granularity for movie reviews might be across all reviews regardless of genre. The features that predict a positive or negative review do not seem to be linked to movie genres, with the slight exception of horror reviews.

Web scraping can be used to build large natural language corpora of user-generated texts. With enough flexibility, these corpora can be organized and labelled to investigate the problems of domain definition. In my analysis, I looked at genres of movies. However, genre labels are deceptive in their simplicity. Sites like IMDb allow multiple genre tags for single titles, and it is common for a movie to sit somewhere in the middle of several genre definitions.

There are many ways to more specifically define subdomains of movie reviews that might make interesting investigations. For example, one might split up reviews by year of release or by whether it was a mainstream or independent release. By looking at movie genres in more detail, it would be possible to explore the challenges of multiple overlapping categories for movies.

Movie genres are just one possible sub-category of movie reviews, which are themselves a small subset of potential sentiment analysis data. It would be an interesting task to test the cross-domain generalizability of reviews for different products on Amazon or from posts on one social media site to posts from another. It is still unclear how strictly domains should be defined for sentiment analysis applications. However, my results show that, at least for movie genres, there are slight differences in performance on the level of specific genres.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes, and Blenders: Domain Adaptation for Sentiment Classification.

Andrew Maas, Raymond Daly, Peter Pham, Dan Huang, Andrew Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics.*

Bo Pang and Lillian Lee. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the Association for Computational Linguistics*.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. 1-135.

Dai Quoc Nguyen, Dat Quoc Nguyen, Thanh Vu, and Son Bao Pham. 2014. Sentiment Classification on Polarity Reviews: An Empirical Study Using Rating-based Features.

“Genre Definitions.” 2018. Internet Movie Database. Accessed April, 2018. https://help.imdb.com/article/contribution/titles/genres/GZDRMS6R742JRGAG

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD-2004). Aug 22-25, 2004, Seattle, Washington, USA.

Appendix. The Scraping Process (Step by Step)

genreIDs -> titles:

‘Horror’ : ‘Jaws 2’

titles -> IMDb IDs

‘Jaws 2’ : ‘tt0077766’

IMDb IDs -> urls (for scraping)

‘tt0077766’ : ‘http://www.imdb.com/title/tt0077766/reviews/_ajax’

url -> review content

‘http://www.imdb.com/title/tt0077766/reviews/_ajax’ :

tt0077766.txt

http://www.imdb.com/title/tt0077766/reviews/_ajax

Jaws 2 tt0077766,

Review Count: 267

Title: As far as sequels go, this one deserves another bite!

Date: 13 April 2003

Rating: 7

When Jaws was released in 1975, I don't think audiences knew what hit them...

Title: Pacing could have been more tight, but it's often suspenseful and exciting.

Date: 18 May 2001

Rating: N/A

As a sequel to an immensely popular classic, Jaws 2 had a...