Paper 11001-2016

# Analysis of IMDB Reviews For Movies And Television Series using SAS® Enterprise Miner™ and SAS® Sentiment Analysis Studio

Ameya Jadhavar, Prithvi Raj Sirolikar, Dr. Goutam Chakraborty, Oklahoma State University

.

## ABSTRACT

Movie reviews provide crucial information and act as an important factor when deciding whether or not to spend money on seeing a film in the theater. Each review reflects an individual's take on the movie and there are often contrasting reviews for the same movie. Going through each review may create confusion in the mind of a reader. Analyzing all the movie reviews and generating a quick summary that describes the performance, direction, screenplay among other aspects will be helpful to the readers in understanding the sentiments of movie reviewers and to decide if they would like to watch the movie in the theatres.

In this paper, we demonstrate how the SAS® Enterprise Miner nodes enable us to generate a quick summary of the terms and their relationship with each other which describe the various aspects of a movie. The Text Cluster and Text Topic nodes are used to generate groups with similar subjects such as genres of movies, acting and music. We used SAS® Sentiment Analysis Studio to build models that helped us classify 15,000 reviews where each review is a separate document and the reviews were equally split into good or bad. The Smoothed Relative Frequency and Chi square statistical model is found to be the best model with overall precision of 78.37 %. A rule based model is built to explain the rules that are used to classify reviews as good or bad. As soon as the latest reviews are out, such analysis can be performed to help viewers quickly grasp the sentiments of the reviews and decide if the movie is worth watching.

## INTRODUCTION

IMDB is the most popular website for movie ratings and movie reviews. Imagine being able to analyze the reviews and understand what exactly the customers liked or disliked. Using text mining we can find the terms that are most commonly used in the reviews and how it affects the movie reputation. We can analyze each term in the text and see which other terms it is strongly related to. Doing so we can gauge the customer satisfaction or dissatisfaction with the movie which may affect the revenue generated by the movie either positively or negatively. Using sentiment analysis we can build models on the existing reviews and be able to predict the new reviews as good or bad. Sentiment mining can unearth the reasons why the movie would be a hit or a failure. Movie makers can use this analysis to improve the quality of the movies to meet the expectations of the general audience and to generate maximum revenue.

## DATA ACCESS

The data for this research paper contains television series and movie reviews taken from http://ai.stanford.edu/~amaas/data/sentiment/. It contains 25,000 text documents for training and 25,000 for testing. For the purpose of this paper we considered the first 15,000 text documents as the data needed for analysis. It also contains the URL's from where the text has been extracted and saved. There is one URL for a corresponding text document in both training and validation data.

## DATA DICTIONARY

| Variable | Level | Description |
|---|---|---|
| ID | ID | This field represents the unique review number |
| Review | Text | This variable represents the actual movie review posted by a person |

**Table 1: Data Dictionary**

## METHODOLOGY

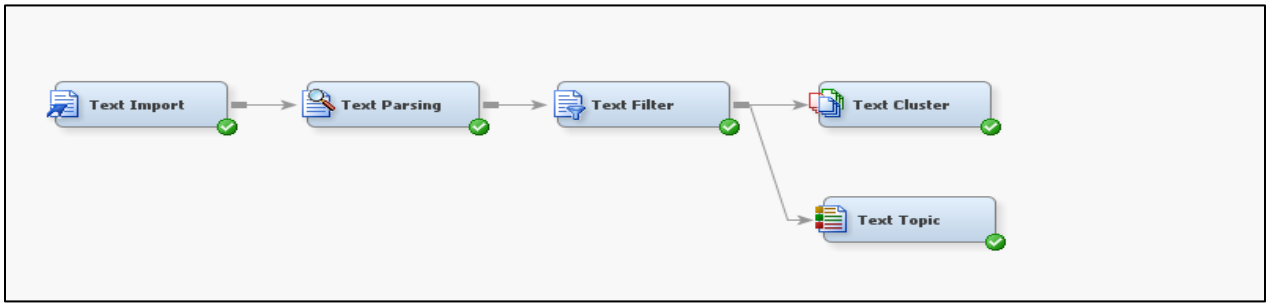



**Figure 1: Text Mining Process**

### Text Import

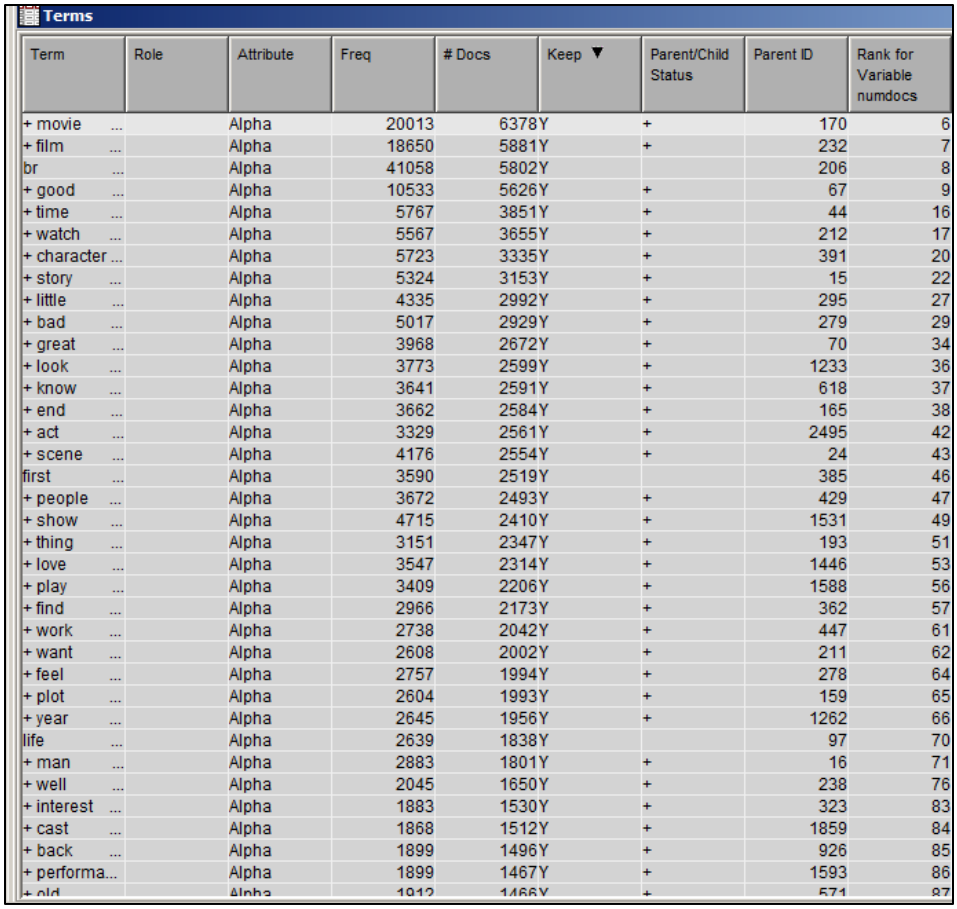
Since the data is available in multiple text documents, it is imported in SAS® Enterprise Miner™ using the text import node. The source points to the folder that contains the reviews in text documents, with one review in one document. The destination points to an empty folder created to write the reviews that are read. The Text Import node converts documents with different formats into text files in the destination folder.

### Text Parsing

After importing the text, the text parsing node is attached to it and a few modifications are made to clean up the unstructured text data. Using the properties panel,

- the 'find entities' option is set to standard,
- the 'detect different parts of speech' option is set to no to be able to represent one word or term as a whole and not have repetitive terms with different parts of speech
- abbr, prop and num parts of speech have been ignored apart from the default options.

The text parsing node also generates the term by frequency document matrix which is used to understand the most frequently occurring term and the number of documents it has occurred in. It is also used to analyze the terms that are rarely used. Ideally the terms that are used moderately are the ones that are the most helpful in exploration and modeling.



Terms

| Term | Role | Attribute | Freq | # Docs | Keep ▼ | Parent/Child Status | Parent ID | Rank for Variable numdocs |
|---|---|---|---|---|---|---|---|---|
| + movie | | Alpha | 20013 | 6378 | Y | + | 170 | 6 |
| + film | | Alpha | 18650 | 5881 | Y | + | 232 | 7 |
| br | | Alpha | 41058 | 5802 | Y | | 206 | 8 |
| + good | | Alpha | 10533 | 5626 | Y | + | 67 | 9 |
| + time | | Alpha | 5767 | 3851 | Y | + | 44 | 16 |
| + watch | | Alpha | 5567 | 3655 | Y | + | 212 | 17 |
| + character | | Alpha | 5723 | 3335 | Y | + | 391 | 20 |
| + story | | Alpha | 5324 | 3153 | Y | + | 15 | 22 |
| + little | | Alpha | 4335 | 2992 | Y | + | 295 | 27 |
| + bad | | Alpha | 5017 | 2929 | Y | + | 279 | 29 |
| + great | | Alpha | 3968 | 2672 | Y | + | 70 | 34 |
| + look | | Alpha | 3773 | 2599 | Y | + | 1233 | 36 |
| + know | | Alpha | 3641 | 2591 | Y | + | 618 | 37 |
| + end | | Alpha | 3662 | 2584 | Y | + | 165 | 38 |
| + act | | Alpha | 3329 | 2561 | Y | + | 2495 | 42 |
| + scene | | Alpha | 4176 | 2554 | Y | + | 24 | 43 |
| first | | Alpha | 3590 | 2519 | Y | | 385 | 46 |
| + people | | Alpha | 3672 | 2493 | Y | + | 429 | 47 |
| + show | | Alpha | 4715 | 2410 | Y | + | 1531 | 49 |
| + thing | | Alpha | 3151 | 2347 | Y | + | 193 | 51 |
| + love | | Alpha | 3547 | 2314 | Y | + | 1446 | 53 |
| + play | | Alpha | 3409 | 2206 | Y | + | 1588 | 56 |
| + find | | Alpha | 2966 | 2173 | Y | + | 362 | 57 |
| + work | | Alpha | 2738 | 2042 | Y | + | 447 | 61 |
| + want | | Alpha | 2608 | 2002 | Y | + | 211 | 62 |
| + feel | | Alpha | 2757 | 1994 | Y | + | 278 | 64 |
| + plot | | Alpha | 2604 | 1993 | Y | + | 159 | 65 |
| + year | | Alpha | 2645 | 1956 | Y | + | 1262 | 66 |
| life | | Alpha | 2639 | 1838 | Y | | 97 | 70 |
| + man | | Alpha | 2883 | 1801 | Y | + | 16 | 71 |
| + well | | Alpha | 2045 | 1650 | Y | + | 238 | 76 |
| + interest | | Alpha | 1883 | 1530 | Y | + | 323 | 83 |
| + cast | | Alpha | 1868 | 1512 | Y | + | 1859 | 84 |
| + back | | Alpha | 1899 | 1496 | Y | + | 926 | 85 |
| + performa... | | Alpha | 1899 | 1467 | Y | + | 1593 | 86 |
| + old | | Alpha | 1912 | 1466 | Y | + | 571 | 87 |

**Figure 2: Text Parsing Output**

The most frequently used terms are movie, film, good, time and watch which makes sense since the reviews are for a movie. The term ‘be’ is misspelt as ‘br’ which is eliminated later using the text filter node.

## Text Filter

The text filter node is added to the text parsing node and is used to eliminate the terms that occur the least number of times in all the documents by manually entering the minimum number of documents it should be present in the properties panel. We can also perform spell check by enabling the option again in the properties panel. Spell check would also suggest the terms that could be potential synonyms. The term ‘thid’ is changed to third, ‘thnik’ to think, ‘cinematograpy’ to ‘cinematography’ and so on.

EMWS1.TextFilter_spellDS

| | Parent # Docs | Term | # Docs | Parent | Role | Parent Role | Min Distance | D |
|---|---|---|---|---|---|---|---|---|
| 51 | 278.0 | easy | 3.0 | easy | PROP_MISC | | 0.0 | |
| 52 | 244.0 | third | 4.0 | third | PROP_MISC | | 0.0 | |
| 53 | 244.0 | thid | 1.0 | third | | | 12.0 | |
| 54 | 2158.0 | thnik | 1.0 | think | | | 10.0 | |
| 55 | 389.0 | cinemtography | 1.0 | cinematography | | | 3.0 | |
| 56 | 389.0 | cinemaphotography | 1.0 | cinematography | | | 10.0 | |
| 57 | 389.0 | cinemaphotography | 2.0 | cinematography | PROP_MISC | | 10.0 | |
| 58 | 389.0 | cinematographed | 1.0 | cinematography | | | 9.0 | |
| 59 | 389.0 | cinematographicly | 1.0 | cinematography | | | 10.0 | |
| 60 | 389.0 | cinematography | 2.0 | cinematography | PROP_MISC | | 0.0 | |
| 61 | 389.0 | cinematograpy | 1.0 | cinematography | | | 3.0 | |
| 62 | 389.0 | cinematopghaphy | 1.0 | cinematography | | | 10.0 | |
| 63 | 389.0 | cinematoraphy | 1.0 | cinematography | | | 3.0 | |
| 64 | 280.0 | future | 1.0 | future | PROP_MISC | | 0.0 | |
| 65 | 2462.0 | great | 4.0 | great | PROP_MISC | | 0.0 | |

**Figure 3: Text Filter Spell Check**

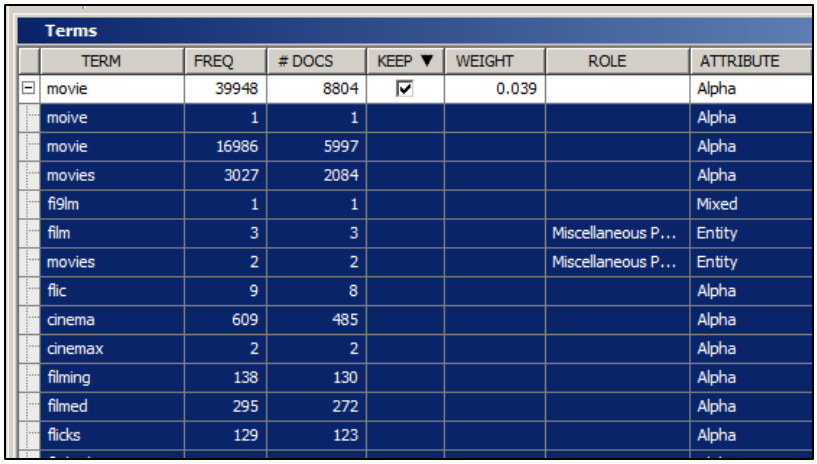
Terms

| Term | Role | Attribute | Status ▲ | Weight | Imported Frequency | Freq | Number of Imported Documents |
|---|---|---|---|---|---|---|---|
| + be ... | | Alpha | Drop | 0.000 | 77440 | 77444 | 9844 |
| + not ... | | Alpha | Drop | 0.000 | 25163 | 25169 | 8121 |
| + have ... | | Alpha | Drop | 0.000 | 18920 | 18933 | 7312 |
| s ... | | Alpha | Drop | 0.000 | 25334 | 25334 | 7276 |
| + do ... | | Alpha | Drop | 0.000 | 17139 | 17149 | 6911 |
| br ... | | Alpha | Drop | 0.000 | 41058 | 41058 | 5802 |
| + see ... | | Alpha | Drop | 0.000 | 9379 | 9406 | 5246 |
| + make ... | | Alpha | Drop | 0.000 | 8929 | 8967 | 5104 |
| + much ... | | Alpha | Drop | 0.000 | 9357 | 9369 | 4984 |
| + just ... | | Alpha | Drop | 0.000 | 7051 | 7054 | 4209 |
| + get ... | | Alpha | Drop | 0.000 | 7069 | 7071 | 4180 |
| + so ... | | Alpha | Drop | 0.000 | 6452 | 6457 | 4019 |
| + go ... | | Alpha | Drop | 0.000 | 5407 | 5426 | 3540 |
| + very ... | | Alpha | Drop | 0.000 | 5561 | 5563 | 3428 |
| + think ... | | Alpha | Drop | 0.000 | 4910 | 4918 | 3273 |
| only ... | | Alpha | Drop | 0.000 | 4334 | 4334 | 3112 |
| + even ... | | Alpha | Drop | 0.000 | 4569 | 4576 | 3109 |
| no ... | | Alpha | Drop | 0.000 | 4768 | 4768 | 3090 |
| + other ... | | Alpha | Drop | 0.000 | 4364 | 4366 | 3034 |
| + little ... | | Alpha | Drop | 0.000 | 4335 | 4344 | 2992 |
| really ... | | Alpha | Drop | 0.000 | 4541 | 4541 | 2979 |
| + that ... | | Alpha | Drop | 0.000 | 4001 | 4002 | 2917 |
| + some ... | | Alpha | Drop | 0.000 | 4179 | 4180 | 2909 |
| + all ... | | Alpha | Drop | 0.000 | 3841 | 3843 | 2842 |
| + like ... | | Alpha | Drop | 0.000 | 3752 | 3752 | 2730 |

**Figure 4: Text Filter Output**

After running the text filter node, we can see that terms such as be, not, have, do are dropped from the text since they do not contribute towards any meaning in the review. Only words that are related to a movie in some way are kept.

Text filter is also used to group synonyms together. It can be done by importing a file with all the synonyms or manually by dragging and dropping the terms into each other.



Terms

| TERM | FREQ | # DOCS | KEEP ▼ | WEIGHT | ROLE | ATTRIBUTE |
|---|---|---|---|---|---|---|
| movie | 39948 | 8804 | ☑ | 0.039 | | Alpha |
| moive | 1 | 1 | | | | Alpha |
| movie | 16986 | 5997 | | | | Alpha |
| movies | 3027 | 2084 | | | | Alpha |
| fi9lm | 1 | 1 | | | | Mixed |
| film | 3 | 3 | | | Miscellaneous P... | Entity |
| movies | 2 | 2 | | | Miscellaneous P... | Entity |
| flic | 9 | 8 | | | | Alpha |
| cinema | 609 | 485 | | | | Alpha |
| cinemax | 2 | 2 | | | | Alpha |
| filming | 138 | 130 | | | | Alpha |
| filmed | 295 | 272 | | | | Alpha |
| flicks | 129 | 123 | | | | Alpha |

**Figure 5: Synonyms Grouping**

The above screenshot shows the synonyms for the term ‘movie’. Terms such as ‘film’, ‘cinema’, ‘flicks’ are grouped together using the interactive filter viewer.

## Concept Links

Concept links can be viewed in the interactive filter viewer from the properties panel of text filter node. It is a type of association analysis between the terms used. Concept links can be created for all the terms that are present in the documents, however it is meaningful to create only for a few important terms. It shows the term to be analyzed in the center and the terms that it is mostly used with as links. The width of the link depicts the strength of association. The wider the link the stronger is the association and the more important it is. Concept links also show how many times the two terms co-exist together in a sentence. A few examples are shown below.

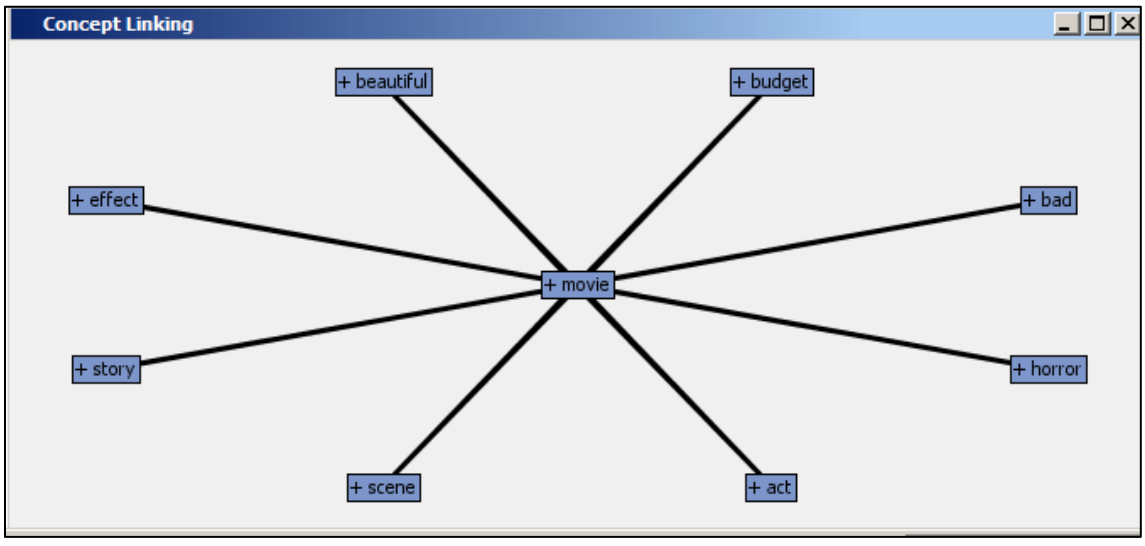



**Figure 6: Concept link for ‘Movie’**

The above concept link is for the term movie. The other terms are related to movie a movie such as a movie would have a budget, it would have a story, a scene or an act, it could be beautiful or bad and so on. Horror, bad and beautiful movies are referenced a lot.

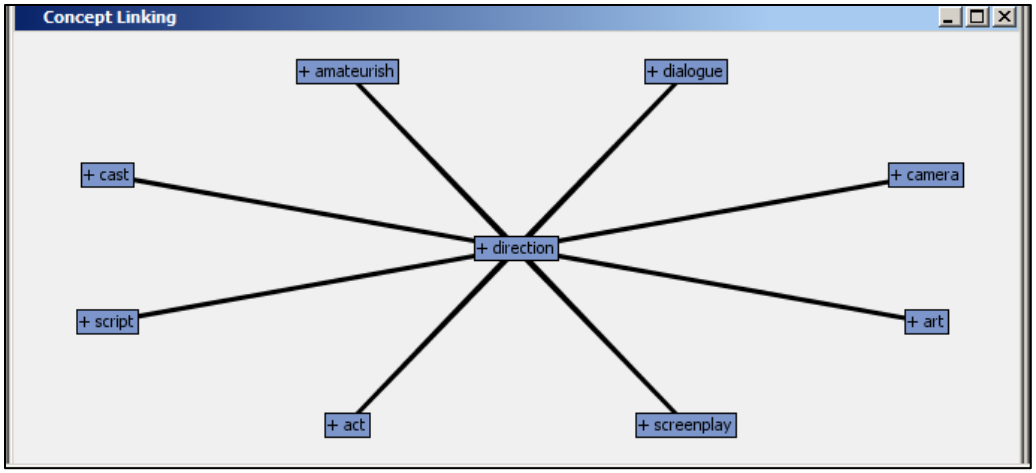



**Figure 7: Concept link for 'Direction'**

Direction is strongly linked with amateurish, indicating that the movies were not good. The other terms such as 'script', 'camera', 'screenplay', and 'dialogue' are always associated with the direction of the movie and thus focusing on those terms would help the movie to be successful.

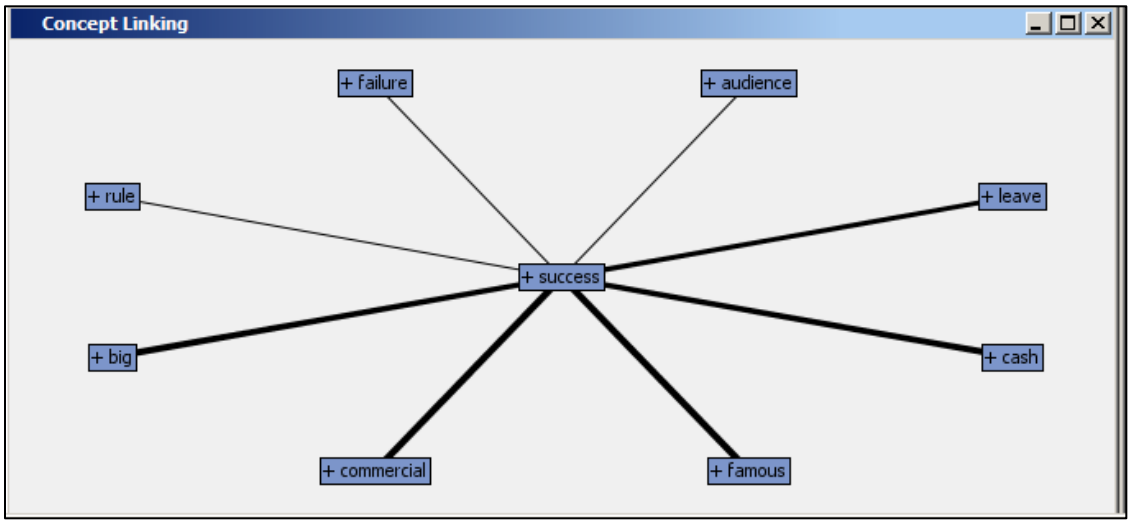



**Figure 8: Concept link for 'Success'**

The terms 'commercial', 'famous' and 'cash' are strongly associated with success which could be indicating that the movie was a commercial success and raked in a lot of cash thereby making the actors in the movie famous. It could also mean that famous actors along with cash rich producers can make the movie a commercial success.

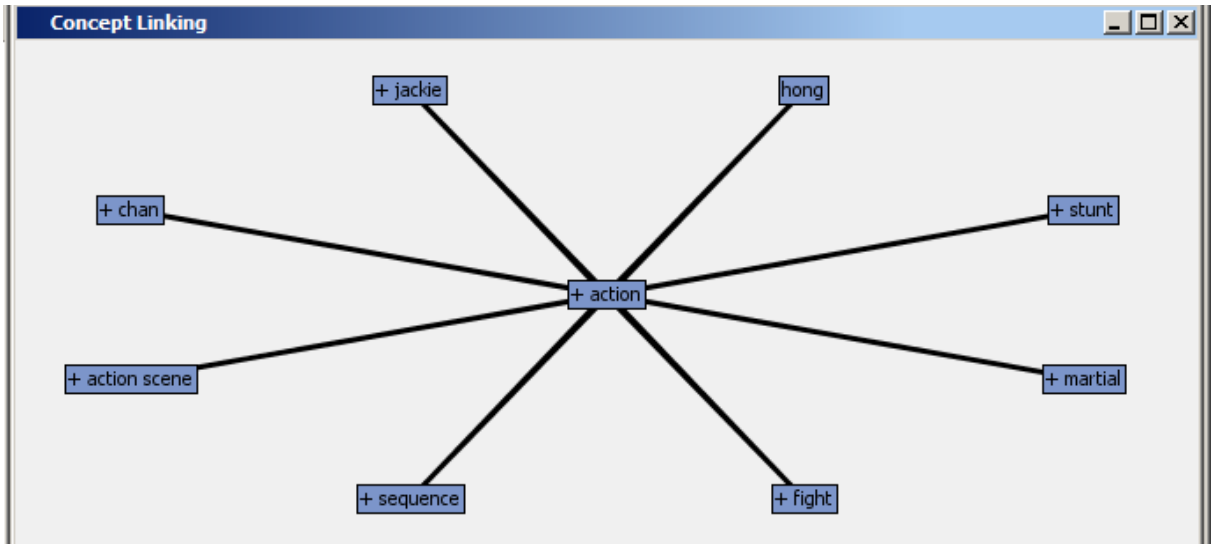



**Figure 9: Concept link for 'Action'**

The term 'action' is associated with the legendary Jackie Chan who is an actor known for his stunts and martial arts. A few reviews mentioned an action scene while some mentioned an entire sequence. Comments on the stunts and fights were made while referring to the action sequences in the movie.

## Text Clustering

Once the text has been filtered using the Text Filter node we group similar terms in the dataset together. SAS® Enterprise Miner™ allows us to group terms closely related to each other into separate clusters of related terms. The properties settings for the Text Cluster Node are set to generate an exact ten cluster solution using Expectation-Maximization Cluster Algorithm and 8 descriptive terms that describe the cluster. The ten clusters generated are well separated from each other as seen in Figure 10.

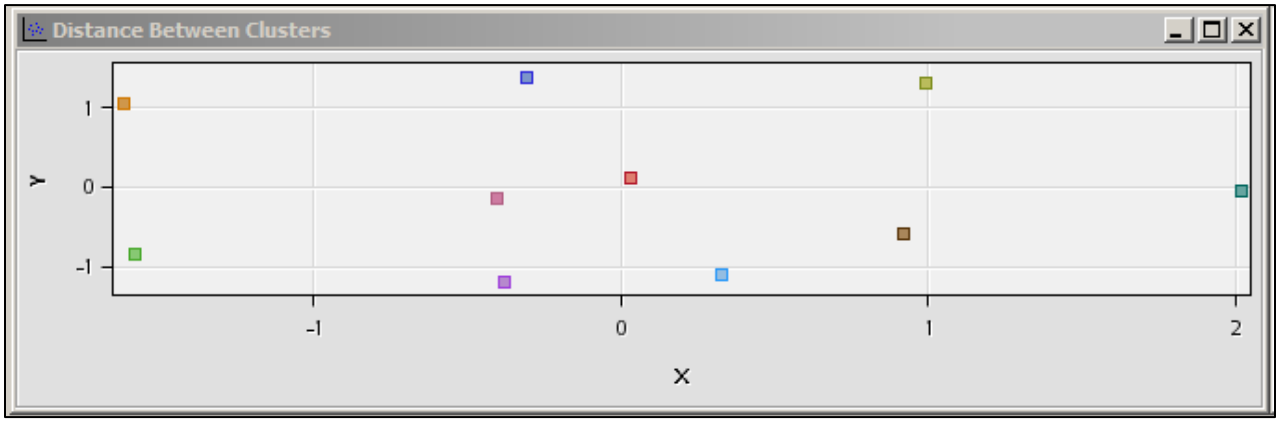



**Figure 10: Distance between Clusters**

The pie chart shows the distribution of the cluster frequencies. Apart from the cluster number two and cluster number eight the frequencies are well distributed among all the clusters as can be seen below.

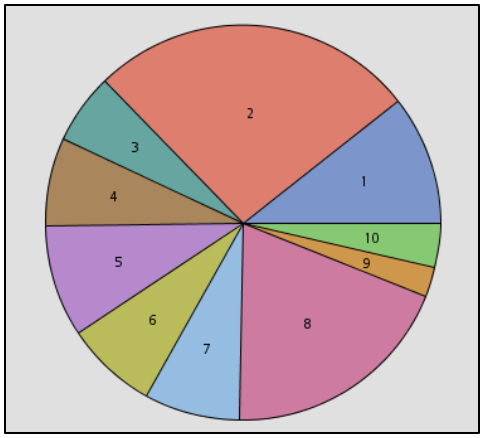



**Figure 11: Distribution of frequencies between clusters**

**Text Clusters Generated**

| Cluster ID | Descriptive Terms | Percentage | Explanation |
|---|---|---|---|
| 1 | +act, +awful, +bad, +money, +movie, +waste, +watch, worst | 11 % | Describes negative reviews for the movie and whether or not it was worth their money |
| 2 | +man, +horror, blood, +live, +play, +scene, +end, +money | 27 % | This cluster is a group of terms for the classification of certain general aspects of the movie. |
| 3 | +episode, +season, +series, +show, +tv, +watch, +funny, +character | 6 % | This cluster clearly groups terms related to television series. |
| 4 | +dance, +musical, +sing, +song, +love, best, great, +cast, | 7 % | This cluster groups all the musical and dance related terms together. |
| 5 | +life, +live, +people, +war, +world, +show, +man, +feel | 9 % | This cluster is a grouping of the attributes that are generally philosophical in nature. |
| 6 | +comedy, +funny, +joke, +laugh, +movie, +watch, +time, +good | 8 % | This cluster is a grouping of the terms that are related to movies that come under the comedy genre. |
| 7 | +book, +love, +movie, +read, great, +act, +best, +watch | 8 % | This cluster groups the terms that determine movies that maybe adopted from books and novels. |

| | | | |
|---|---|---|---|
| 8 | +act, +cast, +character, +role, +play, +work, +feel, +story | 19 % | This cluster groups the terms that are associated with actors in a movie. |
| 9 | +horror, +vampire, +zombie, blood, +bad, +fight, worst, +act | 3 % | This cluster is a grouping of the terms that are related to horror and thriller categories of movies. |
| 10 | +action, +fight, +war, +scene, best, +story, blood, great | 4 % | This cluster is a grouping of the terms that are related to action and war related movies. |

**Table 2: Distribution and Explanation of Text clusters**

### Text Topic

After connecting the Text Filter node in SAS® Enterprise Miner™ we join the Text Topic node which will enable us to combine the term into topics so that we can analyze further. The properties settings for the Text Topic node have been set to generate 7 topics.

| Topic ID | Topic Terms | Explanation |
|---|---|---|
| 1 | +kill, +horror, +killer, +murder, +man | This topic shows the presence of horror and thriller categories of movies in the data. |
| 2 | +bad, +watch, +worst, +movie, +awful | This topic is a group of negative reviews for the movies in the data. |
| 3 | +episode, +show, +series, +season, +tv | This topic quite clearly groups terms related to television series. |
| 4 | +comedy, +role, +funny, +cast, +play | This topic shows the presence of the comedy genre of movies in the data. |
| 5 | +funny, +love, +show, +family, +laugh | This topic is a grouping of the attributes pertaining to light hearted family oriented movies. |
| 6 | +life, +war, +people, +world, live | This topic is a grouping of movies related to war and battle. |
| 7 | +song, +dance, +musical, +sing, +music | This topic groups terms that are related to the songs and dance aspects of a movie. |

**Table 3: Grouping of the Text Topic results**

## SENTIMENT ANALYSIS

SAS® Sentiment Analysis Studio is used to build a statistical model on the text data and to be able to classify the reviews as good or bad using the terms contained in it. The statistical model is run on the test data to check if the model is predicted accurately. 80% of the reviews are used for training the statistical model.

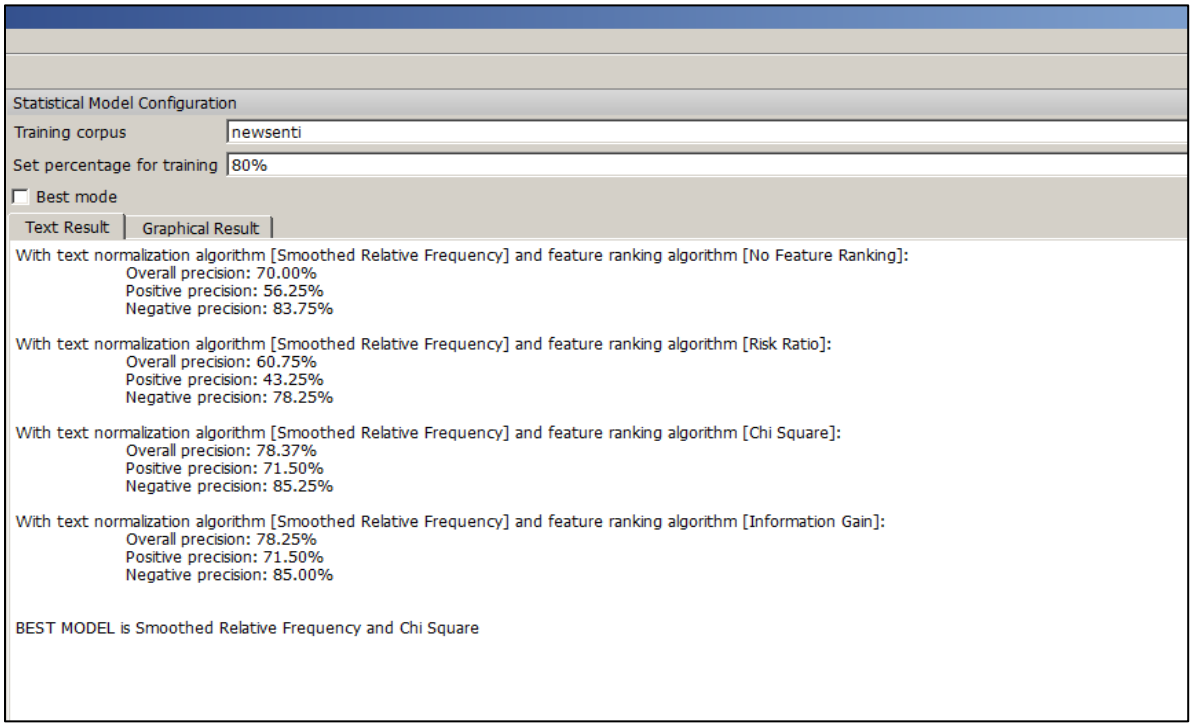



**Figure 12: Statistical Model Output**

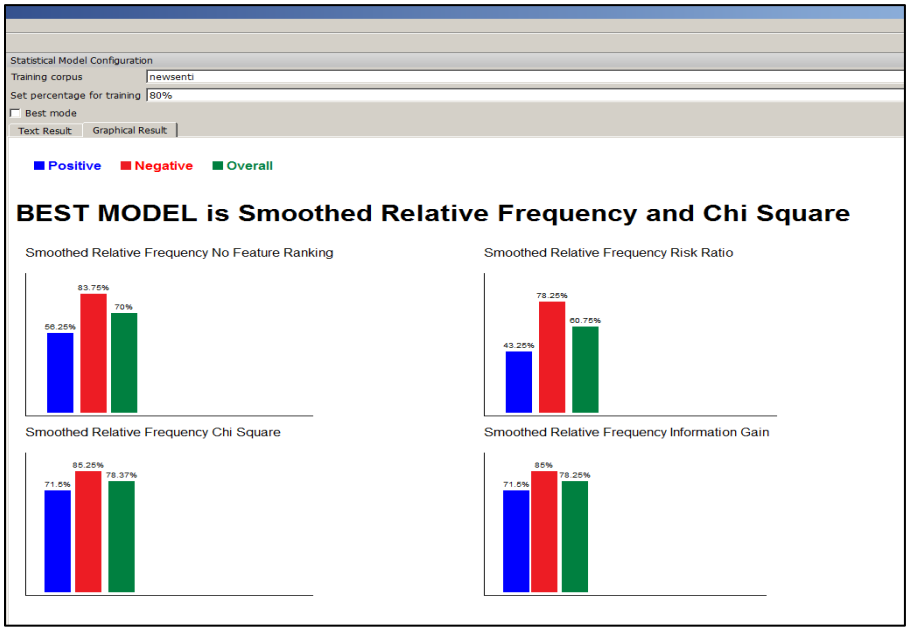



**Figure 13: Statistical Model Graphical Result**

By running the statistical models on the training data we find that the Smoothed Relative Frequency with Chi Square is the best model chosen with its overall precision at 78.37%. The smoothed relative frequency algorithm is a text normalization method that corrects for the length of the document and the number of feature words per document to maintain consistency since some of the documents may be small while others are large. Chi-square is a feature ranking algorithm that basically classifies the features of the document based on its frequency and importance and uses it to build a model.

Now, the model is tested using the test data.

## Model Testing

The statistical model built is used to test the model accuracy on the test dataset for both the positive and negative reviews. We have used a total of 1000 reviews for the testing the accuracy of the statistical model.

### Test for Positive Reviews:

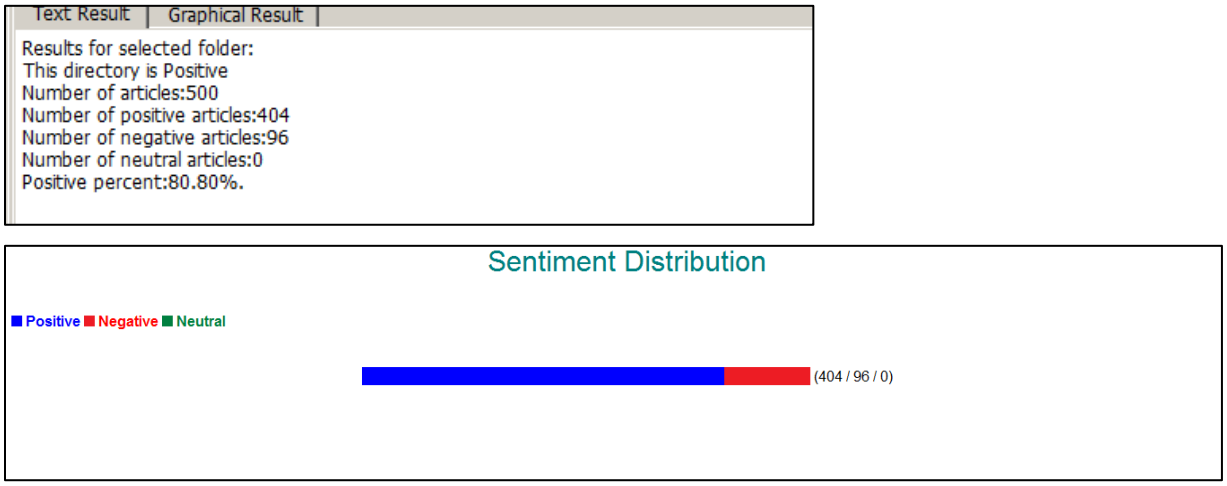



**Figure 14: Positive Reviews Testing Output**

It can be seen that the precision for identifying positive reviews is 80.80%. It is higher than the model accuracy obtained by using the training data. It correctly classifies 404 out of 500 articles as positive.

### Test for Negative Reviews:

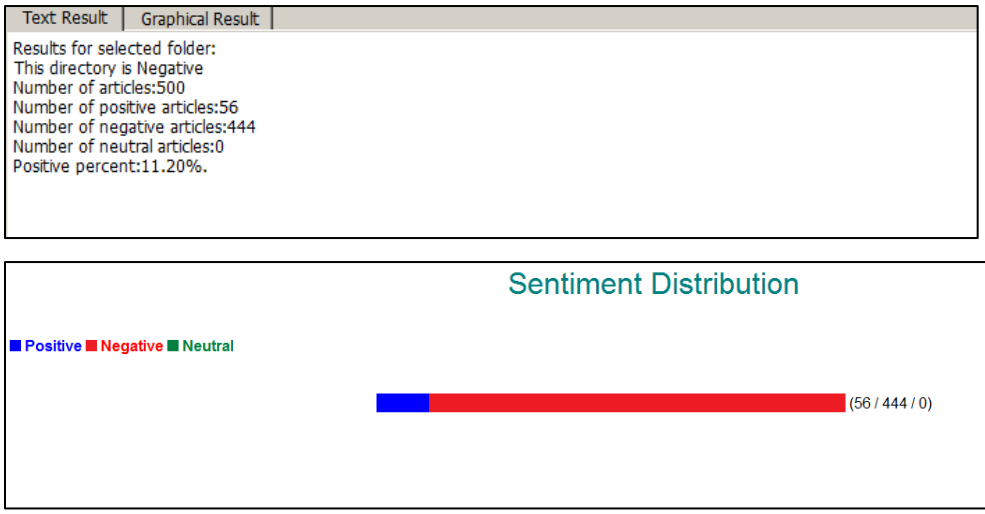



**Figure 15: Negative Reviews Testing Output**

By observing the above table and graph, we can see that the model classifies 88.8% of the documents correctly using the statistical model built which is higher than the percent correctly classified in the training data. 444 out of 500 documents are correctly classified as negative.

Though the statistical model shows good precision in predicting the positive and negative reviews it cannot be explained easily since it is like a black box. We do not know what terms were considered by the model as good or bad. To be able to understand this we have built a rule based model.

## RULE BASED MODEL

### Methodology

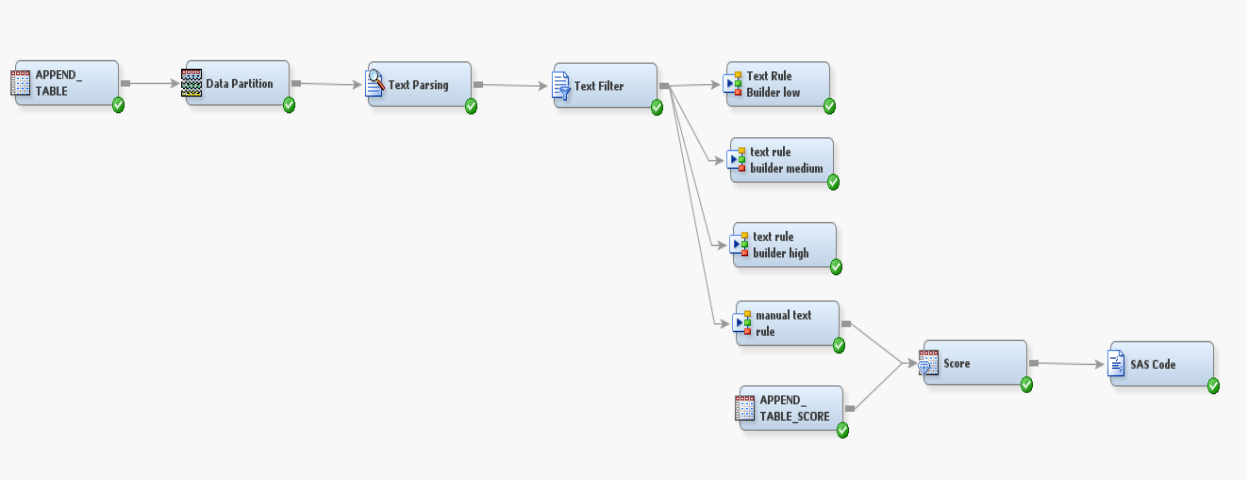



**Figure 16: Rule Based Methodology**

We have a dataset with all the 15,000 reviews and the target variable coded as 1 for ‘positive’ and 0 for ‘negative’. We first use the data partition node to set 70% of the observations as training and the rest 30% as validation. Then the text parsing and text filter nodes are added similar to before. All the properties of the text parsing and text filter node are set the same way as we did before building the clusters. Next we added the text rule builder node with different combination of settings in the properties panel.

The text rule builder node is run with low, medium and high settings for the generalization error, purity of rules and exhaustiveness settings. Amongst these, we found that the text rule builder with the high setting was the best model with the lowest misclassification rate. The misclassification rate for the validation data is 19.92%. To further improve the model accuracy we used the 'change target values' property to manually check if any review was classified incorrectly. An example is shown in figure 17.

| Text | Data Partition | Target Variable | Original Target | Predicted Target | Why Classified | Posterior | Assigned Target |
|---|---|---|---|---|---|---|---|
| I don't want to spend to long here rambling about the plot- you've seen the trailer, and if you haven't its online. I don't recommend seeing it though- it was poorly crafted and didn't pack any of the | Training | Decision | 1 | 0 | lousy | 100.0% | 1 |

**Figure 17: Edited Target Values**

The review clearly shows that it is negative however it was originally classified as positive (1). The model predicted it correctly as negative (0). Hence using our judgement we went ahead and changed the value of the assigned target from positive (1) to negative (0). After making a few more changes the model was run again and now the misclassification rate for the validation data fell to 19.17%. The fit statistics of the model after the manual changes can be seen in figure 18.

Fit Statistics

| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation |
|---|---|---|---|---|---|
| Decision | Decision | ASE | Average Squared ... | 0.033276 | 0.034466 |
| Decision | Decision | DIV | Divisor for ASE | 20998 | 9002 |
| Decision | Decision | MAX | Maximum Absolut... | 0.500367 | 0.499431 |
| Decision | Decision | NOBS | Sum of Frequenci... | 10499 | 4501 |
| Decision | Decision | RASE | Root Average Sq... | 0.182416 | 0.185652 |
| Decision | Decision | SSE | Sum of Squared ... | 698.7209 | 310.2674 |
| Decision | Decision | DISF | Frequency of Cla... | 10499 | 4501 |
| Decision | Decision | MISC | Misclassification ... | 0.166873 | 0.191735 |
| Decision | Decision | WRONG | Number of Wrong... | 1752 | 863 |

**Figure 18: Rule Based Model Fit Statistics**

Now to understand what terms were used to categorize the review as good or bad we will look at the rules that govern them. The rules for positive reviews are seen in figure 19.

| Target Value | Rule # | Rule | True Positive/Total | Precision | Valid True Positive/Total | Valid Precision |
|---|---|---|---|---|---|---|
| 1 | 1 | love & ~bad & ~waste & favorite | 147/155 | 95.45% | 53/56 | 94.64% |
| 1 | 2 | great & ~bad & ~bear & ~waste & performance & ~acting | 257/273 | 95.18% | 129/140 | 93.72% |
| 1 | 3 | wonderfully | 123/133 | 95.08% | 47/55 | 92.86% |
| 1 | 4 | excellent & ~bad & ~mean | 462/513 | 93.77% | 221/257 | 91.03% |
| 1 | 5 | great & ~bad & ~waste & ~bear & love & ~at all | 476/522 | 92.82% | 194/217 | 90.18% |
| 1 | 6 | great & ~bad & ~waste & ~bear & job | 195/214 | 92.91% | 81/93 | 89.51% |
| 1 | 7 | superb & ~bad | 198/215 | 92.99% | 93/106 | 89.28% |
| 1 | 8 | highly & recommend & ~disappoint | 197/211 | 93.19% | 82/90 | 89.56% |
| 1 | 9 | beautifully | 147/167 | 93.02% | 72/83 | 89.59% |
| 1 | 10 | life & ~bad & ~waste & ~annoy & ~decent & ~terrible & ~stupid & discover | 59/62 | 93.06% | 24/27 | 89.39% |
| 1 | 11 | play & ~bad & ~waste & ~awful & ~suppose & ~pointless & ~lousy & ~apparently & ~lame & entertain | 102/111 | 92.91% | 30/36 | 89.08% |
| 1 | 12 | love & ~bad & ~waste & ~terrible & ~plot & ~fail & ~poor & ~lame & always | 164/183 | 92.90% | 70/83 | 88.65% |
| 1 | 13 | wonderful & ~bad | 419/489 | 92.10% | 188/217 | 87.93% |
| 1 | 14 | year & ~bad & ~waste & ~poor & ~terrible & ~awful & ~stupid & ~dull & role | 202/229 | 91.82% | 77/88 | 87.70% |
| 1 | 15 | tremendous | 56/62 | 91.89% | 11/15 | 87.64% |
| 1 | 16 | outstanding & ~bad | 131/143 | 91.92% | 58/67 | 87.58% |
| 1 | 17 | beautiful & ~decent & ~terrible & amaze | 64/65 | 91.97% | 17/19 | 87.59% |
| 1 | 18 | touch & ~suppose & ~bad & true | 48/50 | 92.06% | 14/16 | 87.48% |
| 1 | 19 | favorite & ~bad & ~minute | 345/384 | 91.69% | 153/176 | 87.54% |
| 1 | 20 | late & ~awful & ~stupid & ~bad movie & ~redeem & ~zombie & ~waste & family | 111/129 | 91.54% | 49/64 | 87.46% |

**Figure 19: Positive Classification Rules**

The most important rule containing terms such as love, not bad, not waste and favorite are at the top with a precision of 95.45%. In the validation data, 53 out of 56 reviews containing these terms is correctly classified as positive. The rules are ordered such that the most important rules are written first and the next most important rule is written after that.

The rules for negative reviews are seen in figure 20.

| Target Value | Rule # | Rule | True Positive/Total | Precision | Valid True Positive/Total | Valid Precision |
|---|---|---|---|---|---|---|
| 0 | | 55bad & ~great & bad movie | 305/315 | 96.83% | 136/141 | 96.45% |
| 0 | | 56waste & ~great & ~work | 500/531 | 95.04% | 213/228 | 94.57% |
| 0 | | 57bad & ~great & ~love & horrible | 162/163 | 95.52% | 56/59 | 94.56% |
| 0 | | 58awful & ~love | 433/464 | 95.26% | 191/205 | 94.39% |
| 0 | | 59pointless | 168/182 | 95.01% | 74/79 | 94.42% |
| 0 | | 60bad film | 153/162 | 94.76% | 71/76 | 94.12% |
| 0 | | 61bad & ~great & ~perfect & acting & ~human | 611/651 | 93.93% | 256/279 | 92.97% |
| 0 | | 62poorly | 214/234 | 93.85% | 97/114 | 92.89% |
| 0 | | 63bad & ~excellent & ~wonderful & ~love & ~highly & ~today & ~perfect & ~strong & ~job & poor | 177/180 | 93.90% | 101/104 | 93.03% |
| 0 | | 64terrible & movie | 353/389 | 93.76% | 144/163 | 92.63% |
| 0 | | 65unfunny | 98/108 | 93.81% | 44/46 | 92.78% |
| 0 | | 66atrocious | 68/71 | 93.84% | 24/27 | 92.75% |
| 0 | | 67plot & ~excellent & ~great & ~well & ~enjoyable & ~lovely & ~role & walk | 57/60 | 93.89% | 19/22 | 92.62% |
| 0 | | 68laughable | 135/149 | 93.87% | 63/69 | 92.73% |
| 0 | | 69bear & ~early & ~especially & suppose | 81/85 | 93.86% | 32/35 | 92.66% |
| 0 | | 70bad & ~great & ~relationship & ~fantastic & ~wonderful & ~subtle & ~wife & ~highly & ~human & annoy | 110/113 | 93.88% | 57/59 | 92.82% |
| 0 | | 71script & ~perfect & ~excellent & ~highly & predictable | 47/47 | 93.94% | 16/21 | 92.50% |
| 0 | | 72lame | 258/295 | 93.58% | 93/105 | 92.25% |
| 0 | | 73bad & ~great & ~relationship & ~fantastic & ~wonderful & ~subtle & ~wife & ~emotion & mess | 89/90 | 93.61% | 41/42 | 92.23% |
| 0 | | 74look & ~excellent & ~perfect & ~love & ~young & ~favorite & ~great job & ~fine & ~enjoy & ~life & ~mu... | 110/122 | 93.46% | 48/58 | 91.92% |
| 0 | | 75bad & ~definitely & ~favorite & ~excellent & ~today & stupid | 227/243 | 93.28% | 115/122 | 91.75% |

**Figure 20: Negative Classification Rules**

The most important rule to classify reviews as bad are the terms bad, not great and bad movie with a precision of 96.83%. The model correctly classified 136 out of 141 reviews in the validation data as negative. The other rules are in the order of importance following the first rule.

The statistical model has an overall precision of 78.37% whereas the rule based model has a higher precision of 80.83% on their respective validation data sets.

Now we will use the model built to score the data with 1000 observations having 500 positive and 500 negative. The data used to score already has a target variable coded as '1' for positive and '0' for negative. This can be used to check how many positive and negative reviews were correctly scored by the model. The output of the score node is shown in figure 21.

Class Variable Summary Statistics

Data Role=SCORE Output Type=CLASSIFICATION

| Variable | Numeric Value | Formatted Value | Frequency Count | Percent |
|---|---|---|---|---|
| I_Decision | . | 0 | 407 | 40.7 |
| I_Decision | . | 1 | 593 | 59.3 |

**Figure 21: Scored Data Output**

407 observations are classified as negative (0) and 593 are classified as positive (1). By doing a cross tab of the scored target variable with the actual target variable we can get the percentage of reviews correctly classified.

Upon running the code we can see the output as in figure 23.

```
The FREQ Procedure

Table of decision_original by EM_CLASSIFICATION

decision_original(decision_original)
          EM_CLASSIFICATION(Prediction for Decision)

Frequency|
Percent  |
Row Pct  |
Col Pct  |0        |1        |  Total
---------+---------+---------+
       0 |     351 |     149 |    500
         |   35.10 |   14.90 |  50.00
         |   70.20 |   29.80 |
         |   86.24 |   25.13 |
---------+---------+---------+
       1 |      56 |     444 |    500
         |    5.60 |   44.40 |  50.00
         |   11.20 |   88.80 |
         |   13.76 |   74.87 |
---------+---------+---------+
Total          407       593     1000
              40.70     59.30   100.00
```

**Figure 23: Cross-Tab Output**

Adding the true positive and true negative values we get 795 (351+444) cases correctly predicted out of a total of 1000 cases giving an accuracy of 79.5%

## CONCLUSION

Around 80 percent of the data that is available in the real world is unstructured, of which text data is a major portion. Movie reviews play an important role in determining the popularity level of the movie among the audience and to specifically understand what they liked or disliked in the movie. It also gives insights into what the people expected from the movie and what they actually want to see in it. This information can be leveraged by the movie makers to make better quality movies to cater to the needs and expectations of the people. Concept links can be used to understand the association of one term with others depending on how often they are used together. For example, the term ‘Director’ is strongly associated with ‘screenplay’, ‘script’, ‘camera’ etc indicating that the audience pays more attention to the little nuances of making a movie, which rests on the shoulder of the director. The raw data needs to be parsed and filtered before being analyzed to correct for spelling mistakes, to group synonyms together and to drop the terms that do not contribute in making sense of the data. In general, movie reviews will be made available within hours if not days of the release of the movie and there will be a section of the audience who would just like to know if the movie is worth the money and a get a quick overview of the movie. Using the statistical model and text rule based model built we can identify terms and rules that help us classify reviews into good or bad. This can help people who want value for their money to then decide whether they want to watch the movie or not.

## REFERENCES

1) Learning Word Vectors for Sentiment Analysis by Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

2) Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® by Goutam Chakraborty, Murali Pagolu, Satish Garla.

3) Sentiment Analysis and Opinion Mining by Bing Liu (May 2012).

4) SAS Institute Inc 2014. Getting Started with SAS® Text Miner 13.2. Cary, NC: SAS Institute Inc.

5) Huayi Li, Arjun Mukherjee, Jianfeng Si and Bing Liu. Extracting Verb Expressions Implying Negative Opinions. Proceedings of Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15). 2015.

6) Zhiyuan Chen, Nianzu Ma and Bing Liu. "Lifelong Learning for Sentiment Classification" to appear in Proceedings of the 53st Annual Meeting of the Association for Computational Linguistics (ACL-2013, short paper), 26-31, July 2015, Beijing, China.

## ACKNOWLEDGMENTS

We thank SAS Global Forum 2016 conference committee for giving us an opportunity to present our work. We also thank Dr. Goutam Chakraborty for his continuous support and guidance.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Ameya Ravindra Jadhavar,
Oklahoma State University
Phone: 405-762-2791
Email: ameya.jadhavar@okstate.edu
Ameya Jadhavar is a graduate student enrolled in Management Science and Information Systems at the Spears School of Business, Oklahoma State University. He has been working on healthcare analytics projects with the Center for Health Systems and Innovation, Tulsa as a data analytics intern since June 2015. He is a SAS® Certified Base Programmer, SAS® Certified Statistical Business Analyst and also a SAS® Certified Predictive Modeler. He has the SAS® and OSU Data Mining Certificate and holds the Google Analytics certification. He has co-authored a paper presented at the SCSUG conference in 2014 and had a poster presentation at the SAS® Analytics conference in 2015.

Prithvi Raj Sirolikar,
Oklahoma State University
Phone: 469-412-5991
Email: prithvi.raj.sirolikar@okstate.edu
Prithvi Raj Sirolikar is a graduate student enrolled in Management Science and Information Systems at Spears School of Business, Oklahoma State University. He has two years of work experience working with Serco and Cognizant Technology Solutions in India. He is a SAS® Certified Base Programmer, SAS® Certified Statistical Business Analyst and also a SAS® Certified Predictive Modeler with a SAS® and OSU Data Mining Certificate. He has co-authored a paper presented at the SCSUG 2014 conference and also a poster at the SAS® Analytics conference 2015.

Dr. Goutam Chakraborty
Oklahoma State University
Email: goutam.chakraborty@okstate.edu
Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS and OSU data mining certificate and SAS and OSU marketing analytics certificate at Oklahoma State University. He has published many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has over 25 Years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.