Sentiment Analysis on YouTube Movie Trailer comments to determine the impact on Box-Office Earning

Rishanki Jain, Oklahoma State University

ABSTRACT

The video-sharing website YouTube encourages interaction between its users via the provision of a user comments facility. This was originally envisaged as a way for viewers to provide information about and reactions to videos, but is employed for other communicative purposes including sharing ideas, paying tributes, social networking, and question answering. This study seeks to examine and categorize the types of comments made by YouTube users to understand how the sentiments of users can impact the first day revenue.

INTRODUCTION

In the past, a lot of sentiment analysis work has been done on movie reviews using the IMDB dataset "*Analysis of IMDB Reviews For Movies And Television Series using SAS® Enterprise Miner™ and SAS® Sentiment Analysis Studio*" by **Ameya Jadhavar** . Also, work has been done on YouTube comment scraping and discussed in "*Sentiment Analysis of Movie Review Comments*" by **Kuat Yessenov** to analyze the channels satisfaction using machine learning algorithms like Naive Bayes, Decision Trees, Maximum-Entropy, and K-Means clustering

This study seeks to examine and categorize the types of comments made by YouTube users on popular Hollywood Movie trailers to understand how the sentiments of these users can impact the first day revenue. Also show the trend of box office earnings based on the sentiments after the movie is released. This will help distributors and movie-makers to determine the response rate for the movie in prior by understanding the comments on the trailers and then once the movie gets released the next day earnings can be predicted by looking at the present-day sentiments.

DATA ACCESS

The training data set considered 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 I have considered the first 25,000 text documents as the data needed for analysis and 1000 comments for validation of the model.
However, for the test purpose I have data of 2 Hollywood official movie trailers belonging to different studios. The dataset contains 4,000 comments of users from YouTube trailers for the movie 'Monster Truck' and around 10,000 comments for the Block Buster hit 'Beauty And The Beast'. The dataset contains the user ID, comment, likes on that comment, replies on that comment and the timestamp of the comment.

.

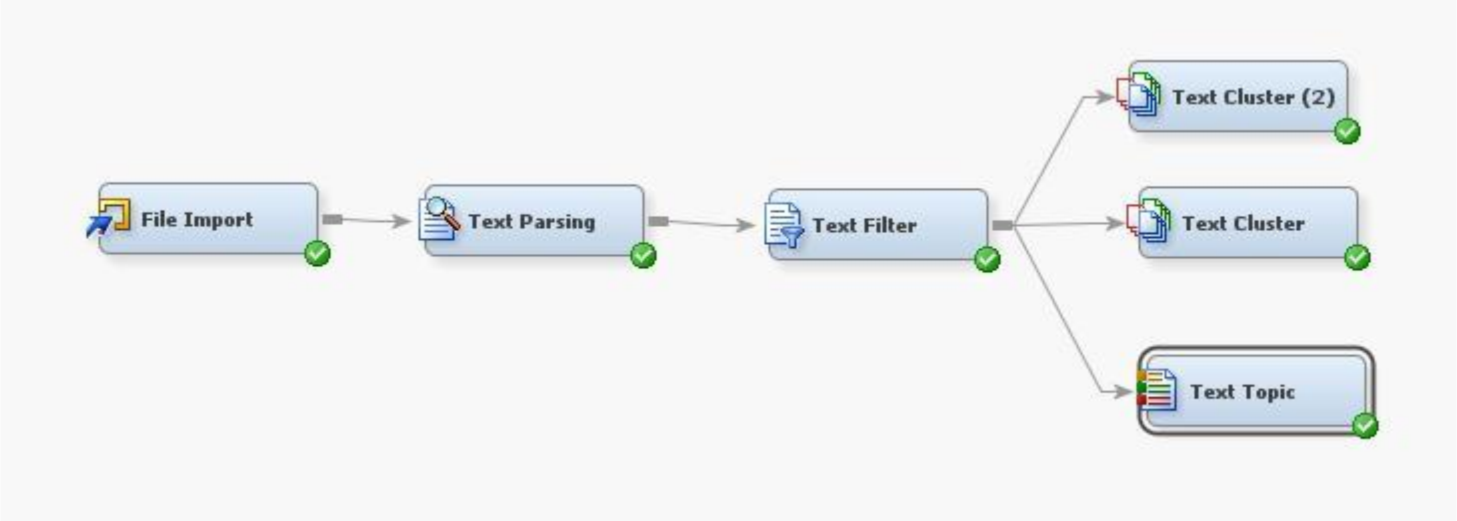
METHODOLOGY

The study has been divided into 3 stages

I. **STAGE 1: TEXT MINING**-

Using the SAS EM Text Miner software, I have first done basic text mining of the comments using the text parsing, text filter, text cluster and text topic nodes. The results tell us about the frequently occurring words and important topics on which broadly the comments are made e.g.: Actors, CGI effects, bad movie, good movie etc.

Using the SAS EM (Text Miner), I have come to the following explorations.



1. Text Import

Since the data is available in multiple text documents, it is imported in SAS® Enterprise Miner™ using the text import node.

2. 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.

3. Text Filter

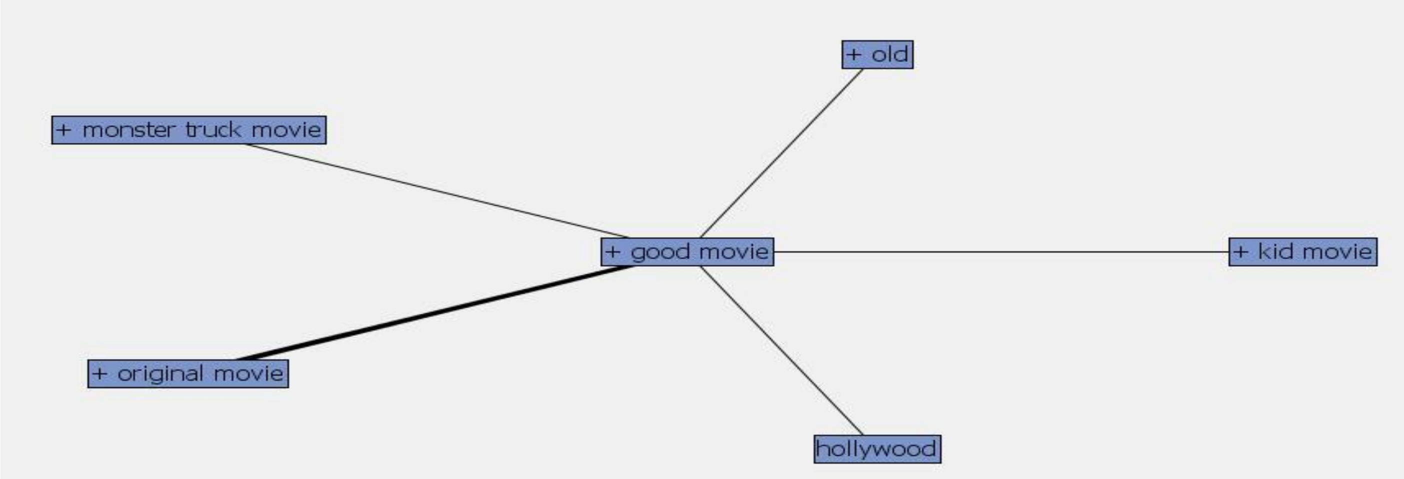
The text filter node is added to the text parsing node and is used to eliminate the terms that occur the least number of times in all the documents by manually entering the minimum number of documents it should be present in the properties panel

4. 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

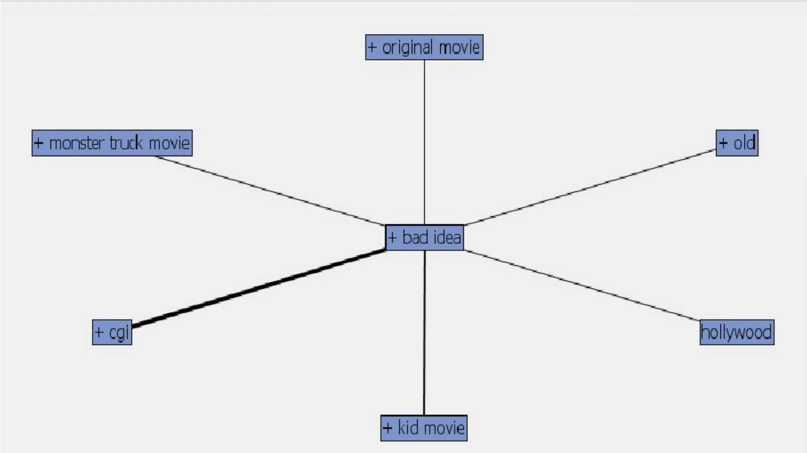
MONSTER TRUCKS

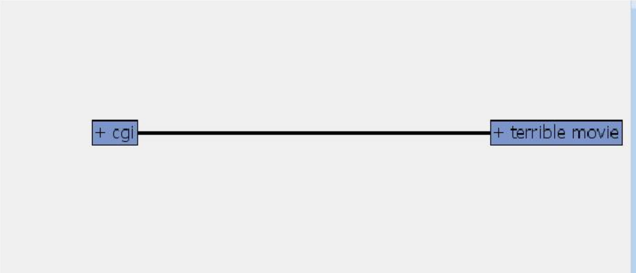
Good movie indicators-



People really liked the original concept of the movie as we can see a high correlation between the good movie and original movie

Bad movie indicators-

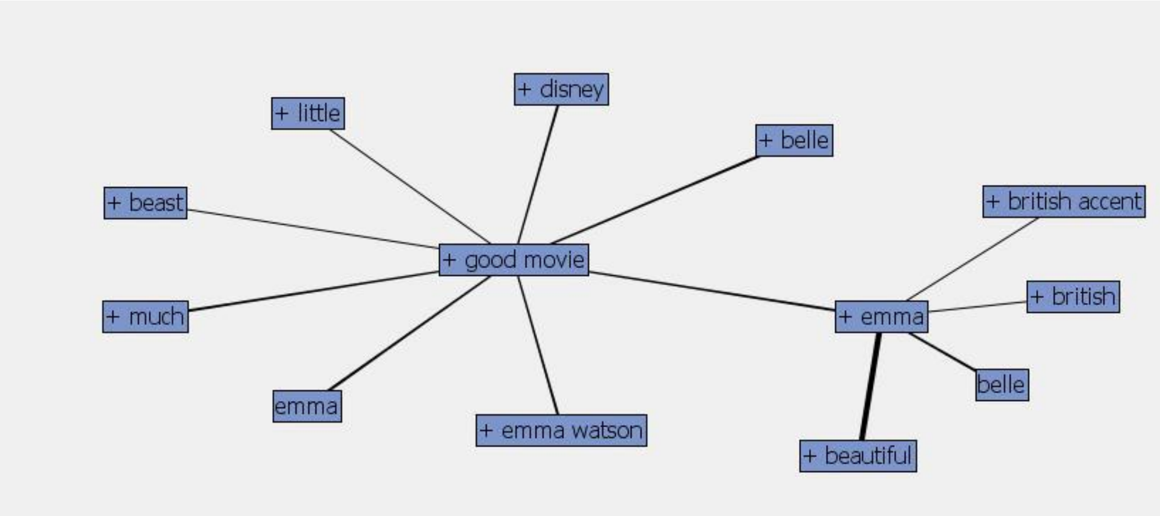




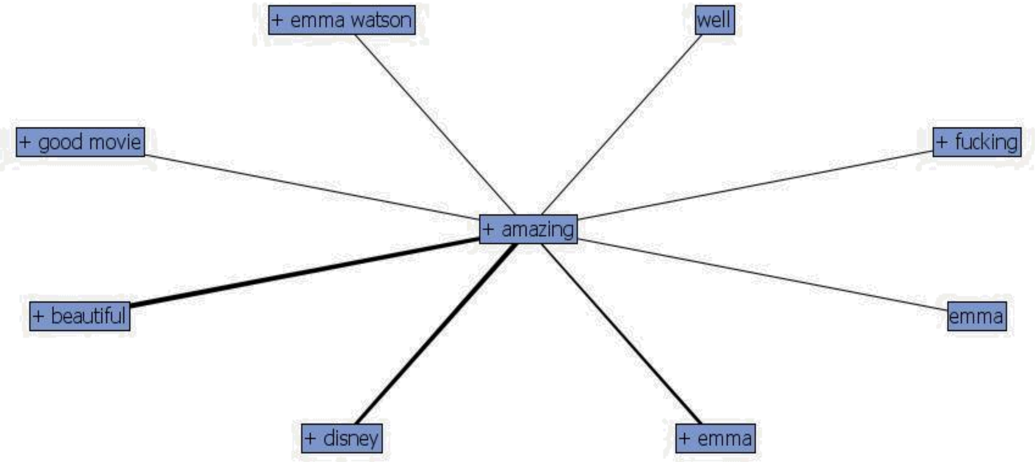
Mostly the CGI effects were not appreciated and highly correlated to the negative comments.

BEAUTY AND THE BEAST

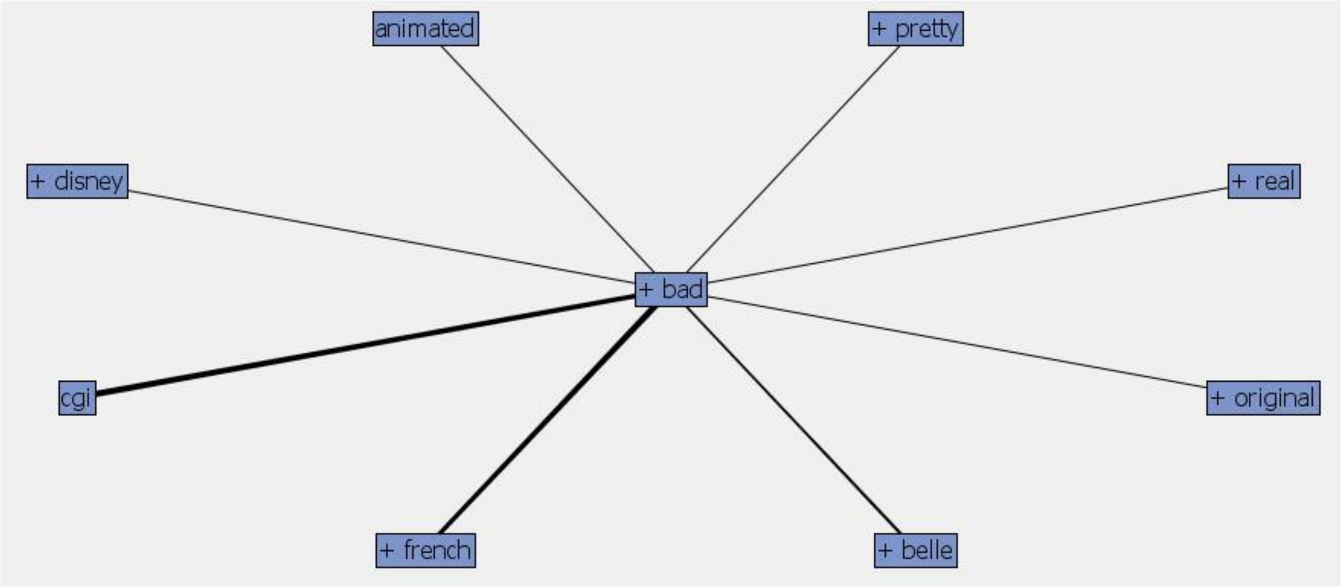
Good movie indicators-



For the positive impacts about the movie people have really appreciated the lead actress- Emma Watson, her character, Belle, and the whole remake of the Disney movie.



Bad movie indicators-



Particularly the viewers were not very happy about the accent used in the movie. They felt that the character was French but the accent used was more of British. Also, there were some disagreement about the CGI effects.

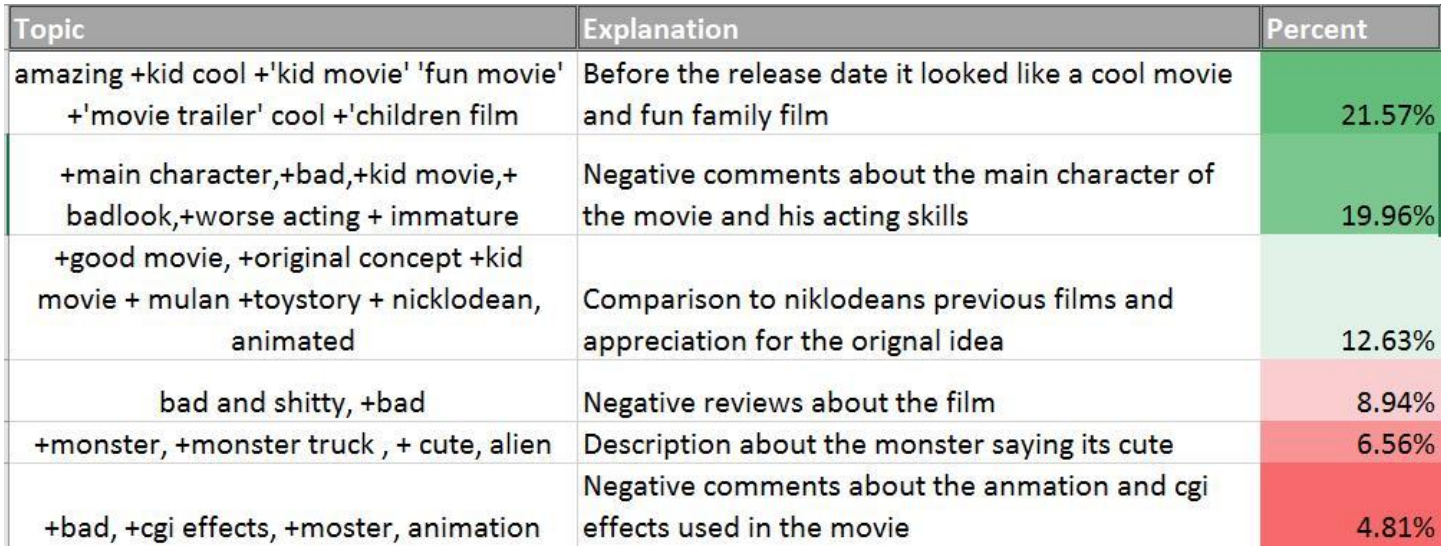
We can further consider this by expanding the terms and understand the reason behind them.

DISCUSSIONS

Major topics and clusters to be considered-

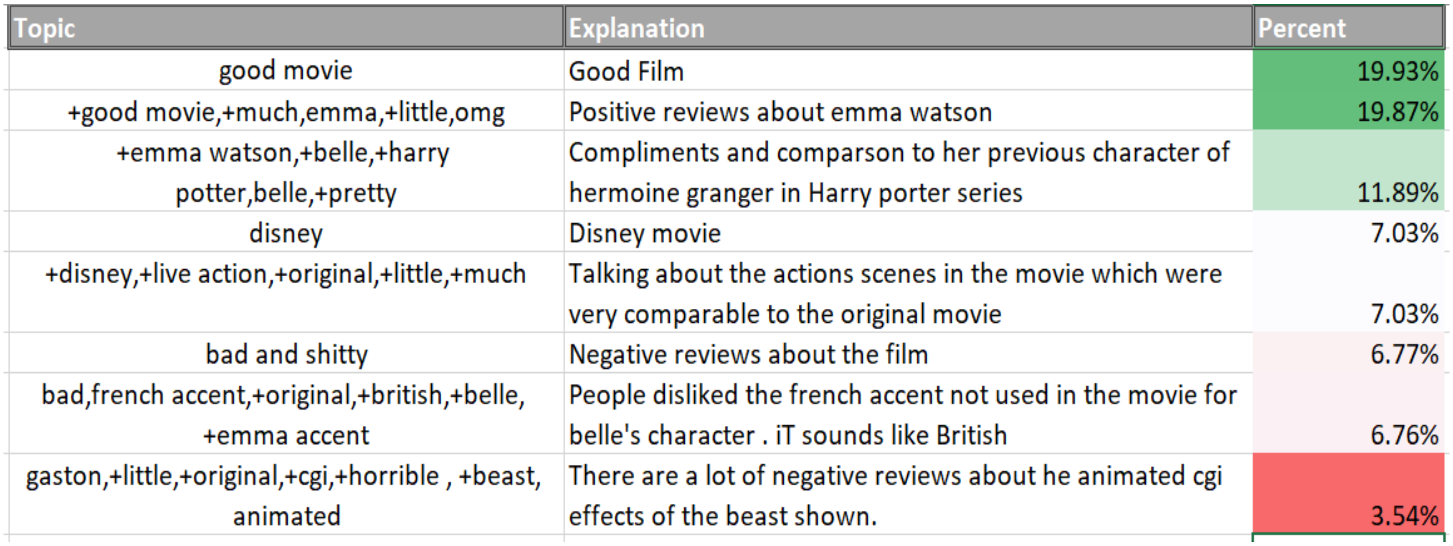
In this section, we have created major topic nodes and clusters from the test that give us an immaculate idea that what were the major attributes people were talking about in the movie.

MONSTER TRUCKS



Topic	Explanation	Percent
amazing +kid cool +'kid movie' 'fun movie' +'movie trailer' cool +'children film	Before the release date it looked like a cool movie and fun family film	21.57%
+main character,+bad,+kid movie,+ badlook,+worse acting + immature	Negative comments about the main character of the movie and his acting skills	19.96%
+good movie, +original concept +kid movie + mulan +toystory + nicklodean, animated	Comparison to niklodeans previous films and appreciation for the orignal idea	12.63%
bad and shitty, +bad	Negative reviews about the film	8.94%
+monster, +monster truck , + cute, alien	Description about the monster saying its cute	6.56%
+bad, +cgi effects, +moster, animation	Negative comments about the anmation and cgi effects used in the movie	4.81%

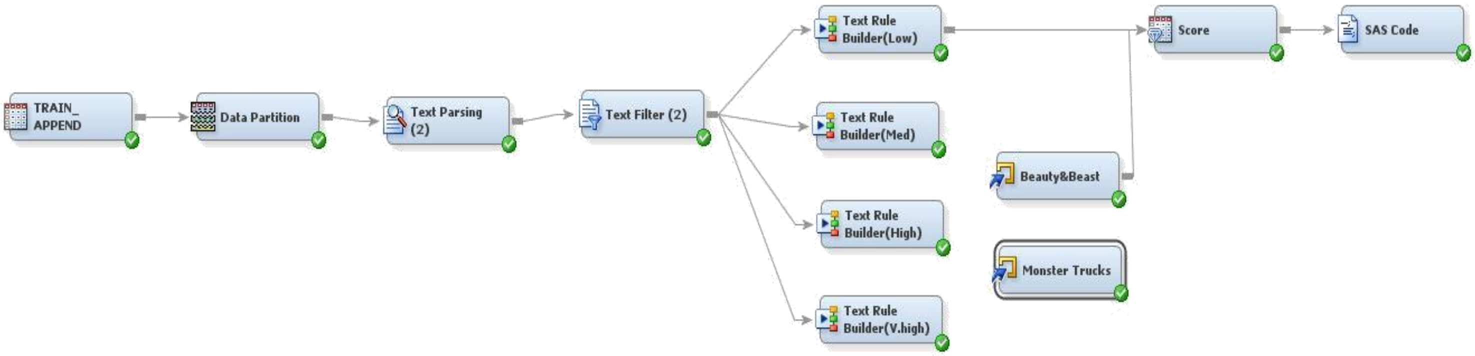
BEAUTY AND THE BEAST



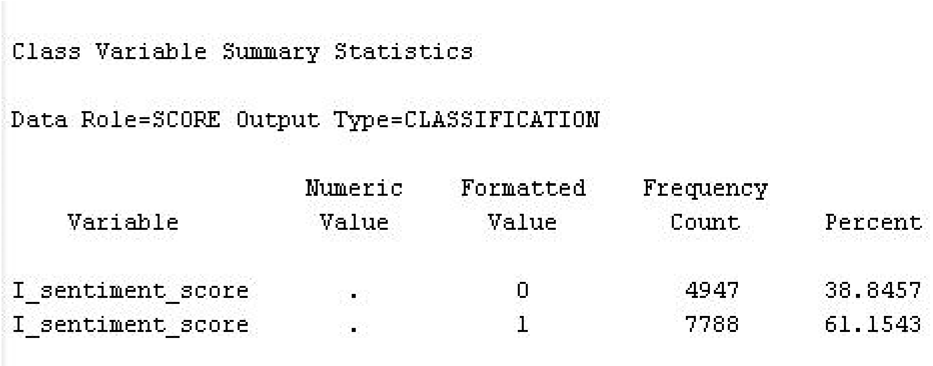
Topic	Explanation	Percent
good movie	Good Film	19.93%
+good movie,+much,emma,+little,omg	Positive reviews about emma watson	19.87%
+emma watson,+belle,+harry potter,belle,+pretty	Compliments and comparson to her previous character of hermoine granger in Harry porter series	11.89%
disney	Disney movie	7.03%
+disney,+live action,+original,+little,+much	Talking about the actions scenes in the movie which were very comparable to the original movie	7.03%
bad and shitty	Negative reviews about the film	6.77%
bad,french accent,+original,+british,+belle, +emma accent	People disliked the french accent not used in the movie for belle's character . iT sounds like British	6.76%
gaston,+little,+original,+cgi,+horrible , +beast, animated	There are a lot of negative reviews about he animated cgi effects of the beast shown.	3.54%

II. STAGE 2: SCORING SENTIMENTS-

Using Prior Training Dataset of IMDB movie reviews in SAS EM Text Miner I did, my text mining obtained the rule builder node extracted features. These rules were then applied on my validation dataset which were again approximately 11000 IMDB polarized comments. Finally, for the testing purpose, I used the YouTube reviews and via the scoring node I got my positive /negative comments for the testing datas



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, I found that the text rule builder with the low setting was the best model with the lowest misclassification rate. The misclassification rate for the validation data is 11.82%. Next when we scored the dataset using the validation dataset we get the following result.



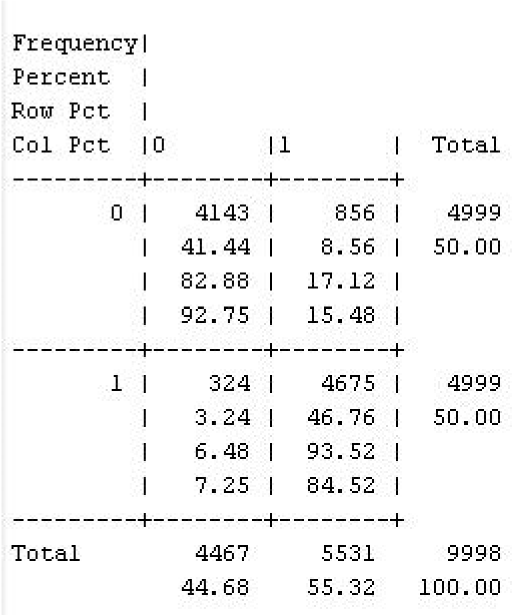
```
Class Variable Summary Statistics

Data Role=SCORE Output Type=CLASSIFICATION

                      Numeric    Formatted    Frequency
    Variable           Value       Value        Count      Percent

I_sentiment_score        .           0           4947      38.8457
I_sentiment_score        .           1           7788      61.1543
```

On creating a cross tab, we find-



```
Frequency|
Percent  |
Row Pct  |
Col Pct  |0       |1       |  Total
---------+--------+--------+
       0 |   4143 |    856 |   4999
         |  41.44 |   8.56 |  50.00
         |  82.88 |  17.12 |
         |  92.75 |  15.48 |
---------+--------+--------+
       1 |    324 |   4675 |   4999
         |   3.24 |  46.76 |  50.00
         |   6.48 |  93.52 |
         |   7.25 |  84.52 |
---------+--------+--------+
Total        4467     5531     9998
            44.68    55.32   100.00
```

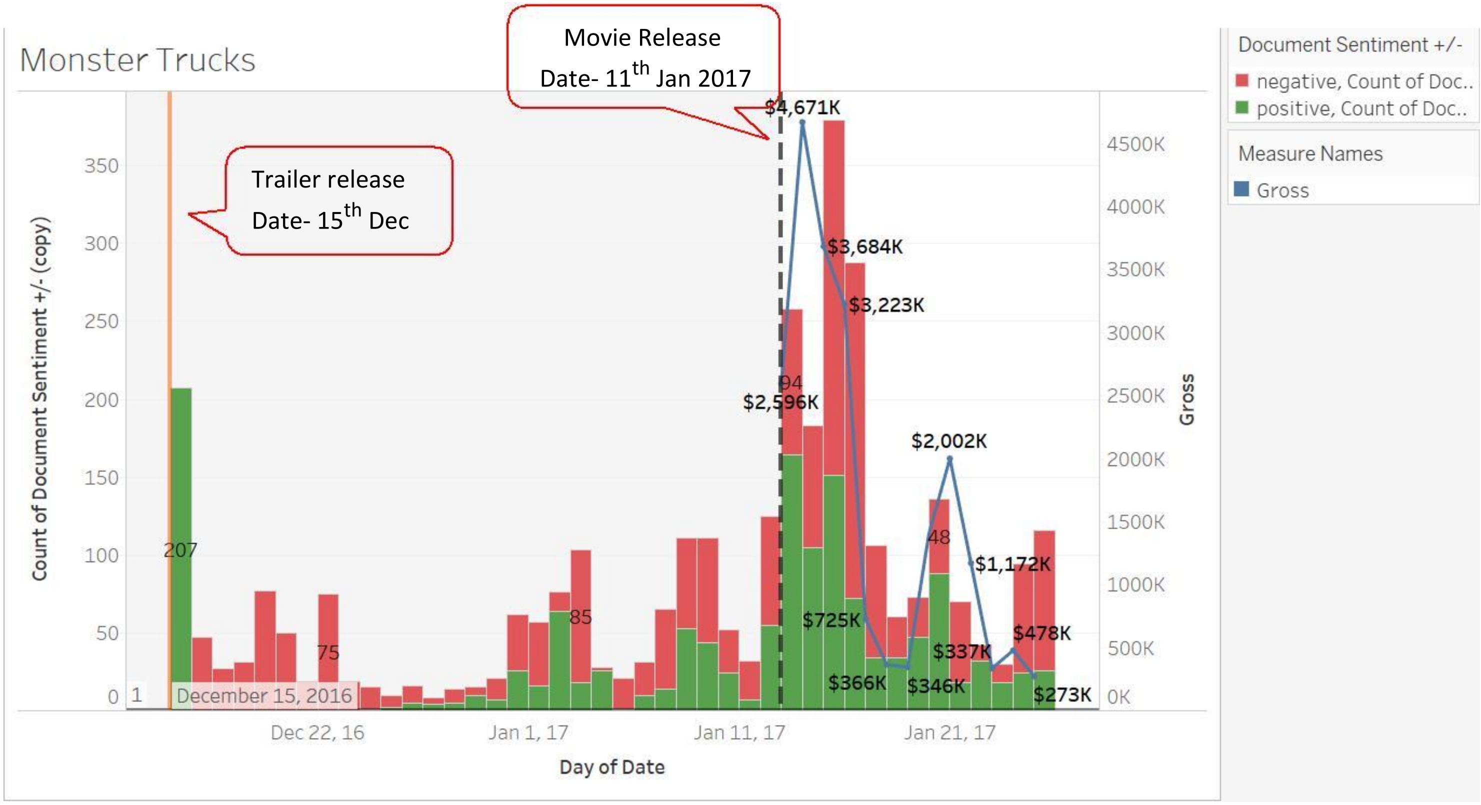
Out of a total 9,998 comments, the model predicted 8,818 (4,143+4,675) comments correctly giving us a prediction accuracy of 88%. Hence, we apply this same model on YouTube data and get our positive /negative comments

III. STAGE 3: GROSS TREND-

Thereafter, using a secondary dataset that contains per day earnings after the movie is released we see a day to day pattern of positive/negative sentiments of people and how does it impact the next day earnings.

MONSTER TRUCKS INSIGHTS

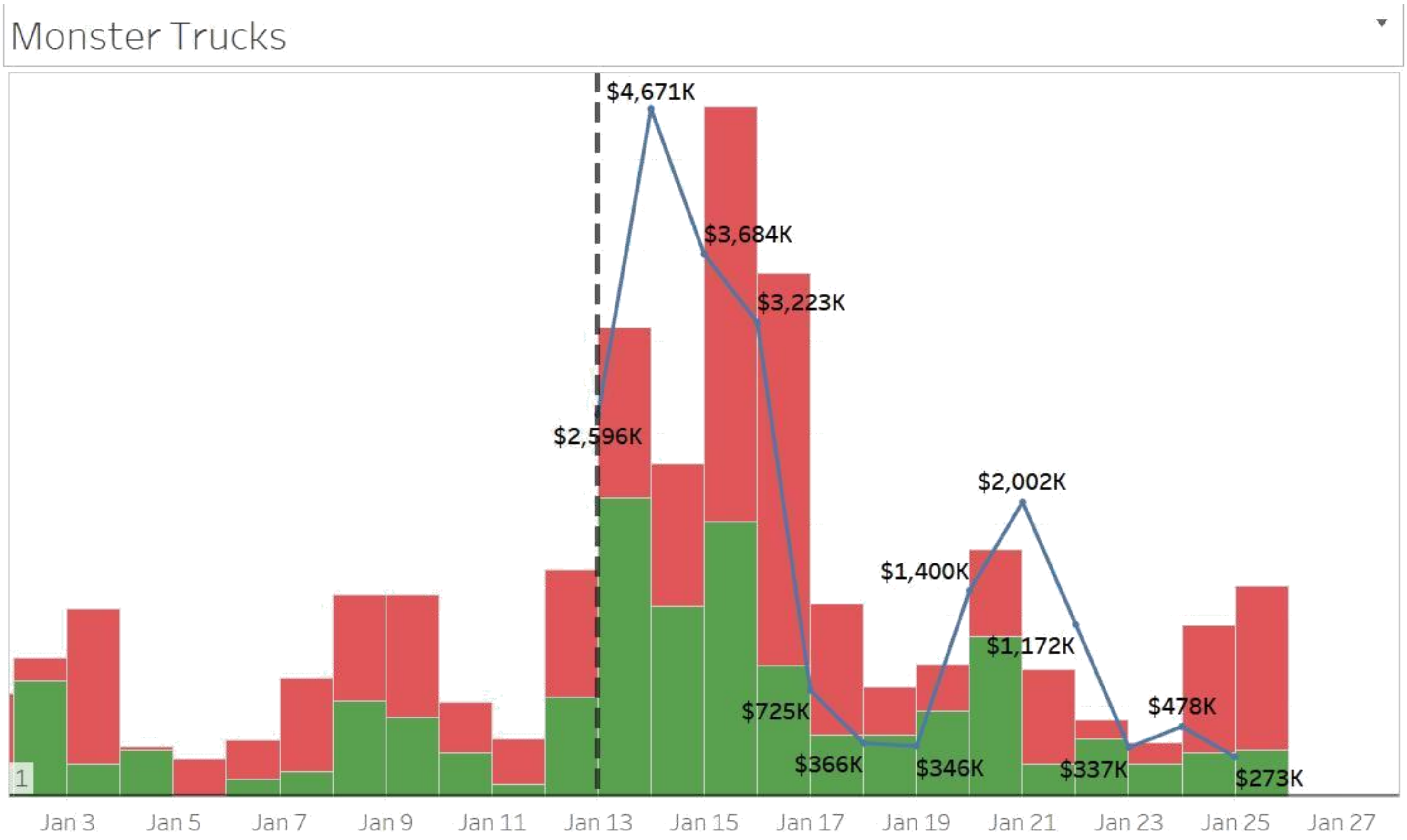
Plotting the positive/negative comments and gross earnings against the Date Variable using Tableau we come across the following graphs



Movie Release Date- 11th Jan 2017

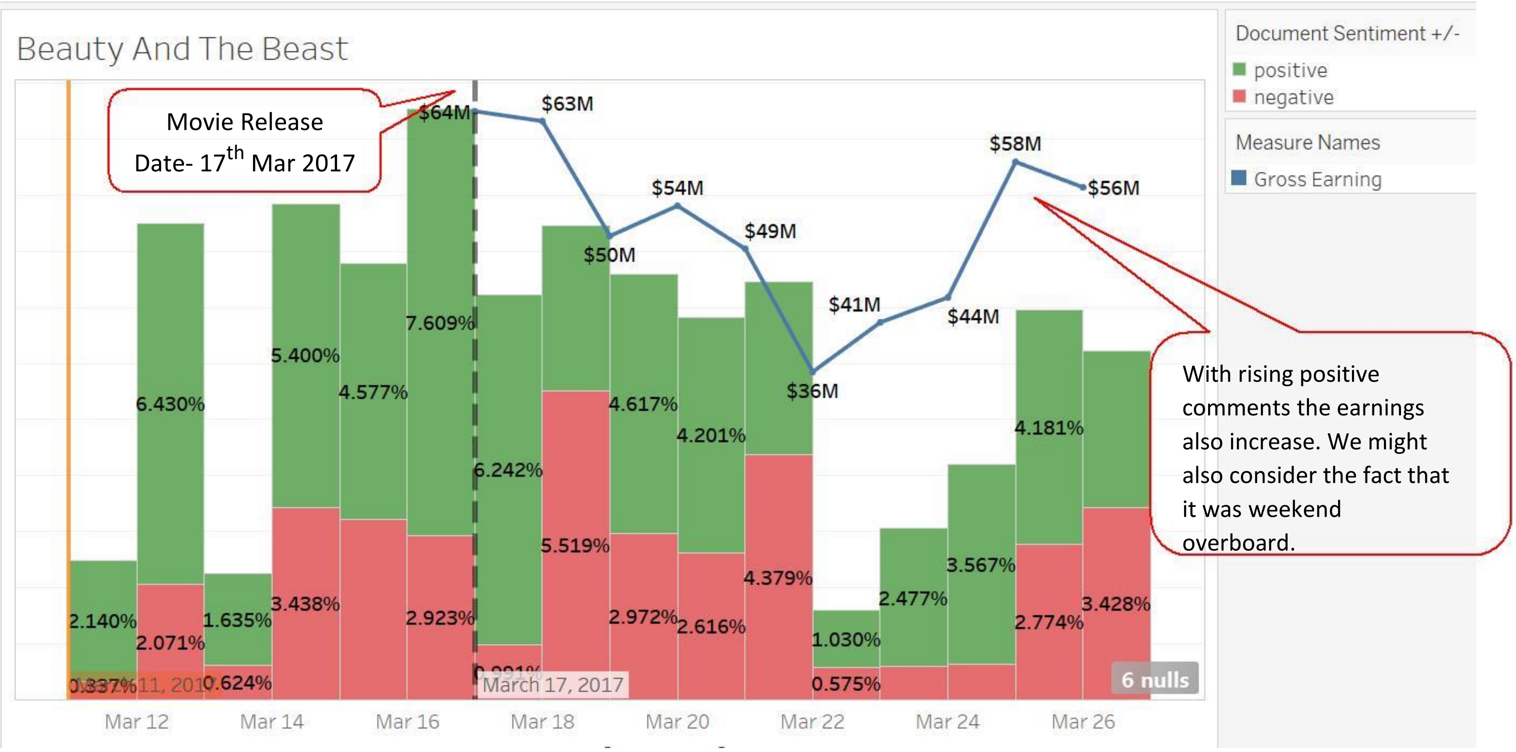
Trailer release Date- 15th Dec

We can see that, once the trailer was released on 15th Dec, there is huge rise in positive comments but, thereafter we can see an increase in negative comments across from 15th Dec to the release date of the movie with little positive comments



A magnified close-up shows us the story after the movie gets released. On January 13th, there is surge of positive comments showing an equivalent high rise in the gross earnings of $4,671K. We can follow how the pattern of the positive comments are impacting the next day earnings. On the 15th-16th of January, we see a high rise of negative comments and subsequently the very next day the earnings fall almost 30%. The same can be seen for 20th, 21st, and 22nd of January. The Earnings are dipping consecutively as the negative comments soar high.

BEAUTY AND THE BEAST INSIGHTS



Movie Release Date- 17th Mar 2017

With rising positive comments the earnings also increase. We might also consider the fact that it was weekend overboard.

On the average, the positive influence in the comments is very high for Beauty and the Beast. There opening day collection was almost 64 million dollars and in this case, surprisingly, we can see that the movie maintains its charm all through the week and does not have major dips, tweaks. The most prominent earning dip is on Mar 22nd where it falls to the maximum low of 36 million dollars. On the previous day, we see approximately 2% rise in our negative sentiments. The movie shows a positive dominance.

CONCLUSIONS

The paper gives us a new insight on how text mining can also be done on YouTube data and give us solutions. We understood the basics of text mining procedure like text parsing, text topic and concept links. It gave us information about what people liked in the movie. For example, in **Monster Trucks**, initially people were looking forward to the release shown by the positive sentiments. However, after the movie launched viewers did not enjoy the CGI effects , the movie script and the acting .There was a sudden increase of negative sentiments along with a drop in the earnings. Similarly, for **Beauty and the Beast** the people were excited about the movie once the trailer launched giving us a wide idea about the first day opening earnings. Also this movie managed to maintain the earnings level and this could be easily verified by there everyday comments on the trailers. Wherever there was a major negative influence, the very next day the earnings fell. Such a conclusion can act like a predictive model for us and tell the movie makers, distributors that how the movie would perform financially the very next day.

FUTURE SCOPE

Right now we have only incorporated revenue of the production house on a daily basis. We would also include the data for no of theatres the movie was launched in, demographics of the audience who went to see it in order to understand the targeted audience and create spontaneous marketing strategies for them. We might also want to include the daily share price impact on the net everyday profit the movie made. In adjunct to sentiment analysis we can create robust models for better information on performance of the movie per day basis.

REFERENCES

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 .

Sharat Dwibhasi, Dheeraj Jami, Shivkanth Lanka, Goutam Chakraborty, 2015, "Analyzing and visualizing the sentiment of the Ebola outbreak via tweets " .

Analysis of IMDB Reviews For Movies And Television Series using SAS® Enterprise Miner™ and SAS® Sentiment Analysis Studio" by Ameya Jadhavar .