***Prediction of Twitter Retweet Count Based on Twitter data***

**Group: Supernova**

**Members: Teena Gorge, Ao Li**

**Qianyi Li, Yuxi Wu, Yaxin Chen**

Content Table

[1. Executive Summary 1](#_Toc513067565)

[2. Business Goal Analysis 1](#_Toc513067566)

[3. Dataset Description 2](#_Toc513067567)

[1) Movie Dataset 2](#_Toc513067568)

[2) Twitter Dataset 3](#_Toc513067569)

[3) Other datasets 3](#_Toc513067570)

[4. System Design 5](#_Toc513067571)

[5. System Implementation 6](#_Toc513067572)

[1) Data Collection 6](#_Toc513067573)

[2) Data Processing 6](#_Toc513067574)

[3) Descriptive analytics 9](#_Toc513067575)

[4) Sentiment Analysis 13](#_Toc513067576)

[5) System Implementation - Build Prediction Model 14](#_Toc513067577)

[a) Neural network 14](#_Toc513067578)

[b) CHAID 15](#_Toc513067579)

[c) Clustering 16](#_Toc513067580)

[6. Conclusions 17](#_Toc513067581)

[7. Limitations 18](#_Toc513067582)

[References 19](#_Toc513067583)

[Appendix I 20](#_Toc513067584)

[Appendix II 21](#_Toc513067585)

# 

# Executive Summary

We collect movie reviews from twitter, weeks before and after the release of movies and calculate the sentiment score for each movie. We analyze various variables from Twitter such as, retweet counts, followers count, and other variables related to a movie such as, budget, box office collection, actor rating, director rating etc. We try to identify the correlation among various variables. Also, we try to predict the retweet count using three different data mining techniques-decision tree(CHAID), k means clustering and neural net.

We considered 14 movies and collected all the required variables. We calculated the sentiment score of each movie using VADER. The correlations were studied from charts made in Tableau and Excel. We found that there is no specific relationship between the sentiment score and box office collection. Similarly, we failed to identify a correlation between compound score and retweets. Another finding is that the tweets about feeling, film type and actor usually have higher retweet count when it is more positive, while tweets about director have higher retweet when it is more negative.

We can see that a movie with lowest sentiment score has a very high retweet count while another movie with around the same sentiment score has a low retweet count, which could be because when a highly anticipated movie gets a bad review, it tends to spread very quickly. This shows that still word of mouth(retweets) is a popular method for promoting movies.

In terms of timing of tweets, we observed that tweets posted in the first week after the film released have higher average retweet count and usually have high sentiment score and that in the second week, the retweets dropped down dramatically. However, second week seems to be very important for box office collection as the total compound score and box collection is the largest in this week. The box office collection must have increased in the second week due to large number of retweets in the first week. Therefore, we further analyzed retweets using SPSS Modeler. Ultimately, compared to K-means clustering, Neural Network and Decision Tree, Neural Network performed better with nearly 80% accuracy. As a result, we found that competition factors and weeks are the most important factors to predict retweeted number.

# **Business Goal Analysis**

With the development of social media, Twitter has become one of the leading platforms that people use to post tweets to express their instant feelings about movies. These instant-feeling tweets, if retweeted a lot, can have an enormous influence on movies’ promotion, reputation, and people’s choice eventually.

Although many practitioners and scholars have found good ways to predict retweet count of tweets, most of their studies focus on the retweet count of tweets of a specific Twitter account, such as “Love, Simon” account, rather than tweets about the movie.

With the use of text analysis, we can take two important factors - sentiment score and tweet labeling into prediction process. Sentiment scores translate how people think about a movie into scores that reflect to what extent people like this movie; tweet labeling helps identify tweets’ content and categorize them into different categories. By knowing these two factors, we can much better understand why people would like to retweet some tweets about movies.

Our goal is to build a mathematical model to predict the retweet count of comments about movies of various genres by using tweet sentiment score and category with other independent variables. From the model, we can know how important a role tweet sentiment score and tweet category are playing in predicting the retweet count. However, we can also see that tweet sentiment score and tweet category alone are not able to fully explain the prediction, that’s why we are introducing other variables to see what other factors are also significant in predicting retweet account.

# Dataset Description

## Movie Dataset

**Table 1 - Movie Data**

|  |  |  |
| --- | --- | --- |
| **Data Description** | | **Count** |
| **Releasing Time** | From March 9th to April 20st | 1 month and a half |
| **Movies** | The Death of Stalin; A Wrinkle in Time; Thoroughbreds; Leaning into the Wind; I Can Only Imagine; Tomb Raider; Love, Simon; Finding Your Feet;  Love After Love, The China Hustle; Outside In; Ready Player One; Chappaquiddick; Pacific Rim Uprising | 14 |
| **Genre** | Action; Drama; Fantasy; Thriller; Documentary; Family; Comedy;  Sci-Fi Action | 8 |

We chose many movies when we started collecting data, but some of them have only a few tweets, so we finally had fourteen movies which covers 8 genres.

## Twitter Dataset

**Table 2 - Twitter Data Description**

|  |  |  |
| --- | --- | --- |
| **Data Description** | | **Count** |
| **Collecting Time** | From March 9th to April 20st | 1 month and a half |
| **Tweets** | The content talking about the 14 movies | 20,633 |
| **Sentiment Score** | Compound sentiment scores of all the movie reviews | 20,633 |
| **Favorites** | number | 20,633 |
| **Retweet Count** | number | 20,633 |
| **Categories** | story, director, film type, actor, ticket, theme, promotion, box office, price, feeling, music, platform | 12 |

To get the tweets we need for the prediction, we used API Tweepy, the open-source - to crawl tweets about fourteen ongoing movies week by week since Tweepy limits us to only get data within one week. What’s more, we intentionally got tweets one week before the movie was released and then got tweets two weeks after the movie was released to see whether the influence of sentiment score week by week. Also, the first two weeks can be critical for a ongoing movie. In total, we collected 20,633 tweets from Twitter while gathering their Favorites and Retweet Count. For each Twitter, we used Python Sentiment Analyzer to get its compound score, which can represent the general feelings people have when they wrote the tweet.

Further, we usedSPSS Modelerto label every tweet based on their content with a specific category. Besides, each tweet can be labeled with more than one category name to fully reflect the content of the tweet. The categories include Story, Director, Film Type, Actor, Ticket, Theme, Promotion, Box Office, Price, Feeling, Music, Platform. Table 6 in Appendix I shows how we categorize different tweets with different labels. Table 2 shows what datasets we got from Twitter.

## Other datasets

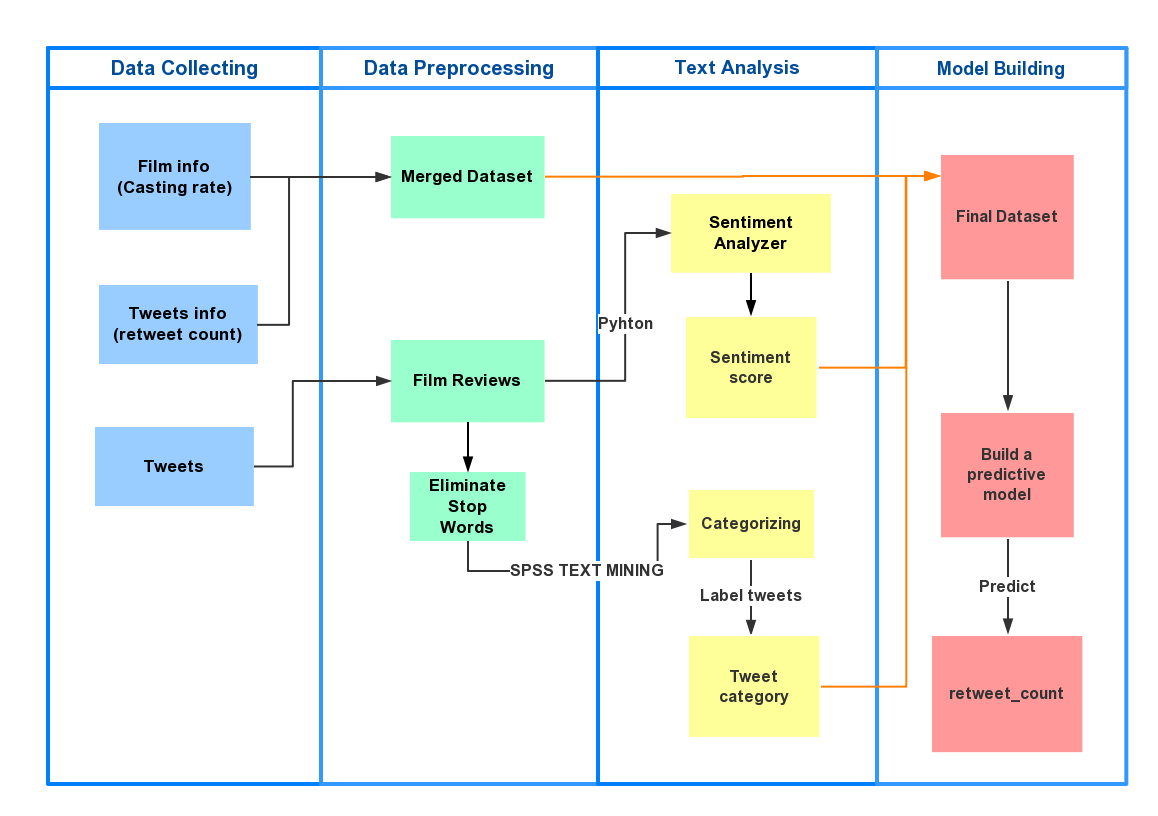
Apart from tweets, we also collected other data to see whether they affect retweet count as Table 3 shows.

**Table 3 - Other Variables Used in Prediction**

|  |  |
| --- | --- |
| **Data/Variable Name** | **Description** |
| **Releasing Time** | The number of week after the movies was released.  “-1” means a week before the movie was released.  “1” means a week after the movie was released.  “2” means two weeks after the movie was released. |
| **Budget** | The amount of money invested in the movie |
| **Box Office** | The entire amount of income of movie ticket sales |
| **Casting Rate** | The average accumulated acting-performance scores of main actors (basically average career score of the first five main casts) from metacritic website |
| **MAPP Rating** | [Motion Picture Association of America](https://en.wikipedia.org/wiki/Motion_Picture_Association_of_America) (MPAA) film rating system, which limits the audience based on their ages.   * **Rated PG**: Parental Guidance Suggested – some material may not be suitable for children * **Rated PG-13**: Parents Strongly Cautioned – some material may be inappropriate for children under 13 * **Rated R**: Restricted – under 17 requires accompanying parent or adult guardian |
| **Director Starmeter Rank** | IMDB Director Starmeter Rank |
| **Competition Factor** | Equals 1 divides the number of the movies released in the same week |
| **Sequel** | Whether the movie is a continued story. N/Y |
| **IMDB Movie Rating** | The rating of the movies from IMDB website. |
| **IMDB Votes** | The votes of the movies from IMDB website. |
| **Rotten Tomatoes (RT) User Rating** | The rating of the movies from Rotten Tomato Website |
| **Rotten Tomatoes (RT) User Votes** | The votes Rotten of the movies from Rotten Tomato Website |
| **Distributor** | The names of the distributor of the movies |

# System Design

For this analysis, we used different tools to reach and present our outcomes. The main tool was Python, which was used to crawl data from Twitter, eliminate stop-words of tweets’ content and use Sentiment Analyzer to get sentiment scores of tweets’ content, and SPSS Modeler, which was used to categorize the content without stop-words. The following diagram depicts the steps taken to complete the analysis.

****

**Figure 1 - System Design**

While we had three datasets for the data we collected, we could process some of the data together. For the 20,633 film reviews we crawled from Twitter, we used them in two aspects. For the first aspect, we used these reviews to get compound sentiment scores; for the second aspect, we eliminated stop-words of all the reviews and categorized them. Thus, we get two variables - Twitter Compound Score and Category for the prediction.

For other data we collected, we combined Releasing Time, Favorite Count, Budget, Casting Rate, MAPP rating, Genre, Director Starmeter Rank, Competition Factor, Sequel, IMDB Movie Rating, IMDB Votes, RT User Rating and RT User Votes Distributor these variables with Twitter Compound Score and Category Model in a merged file. We used all these variables to build a predictive model that hopefully could fully predict the Retweet Count.

# System Implementation

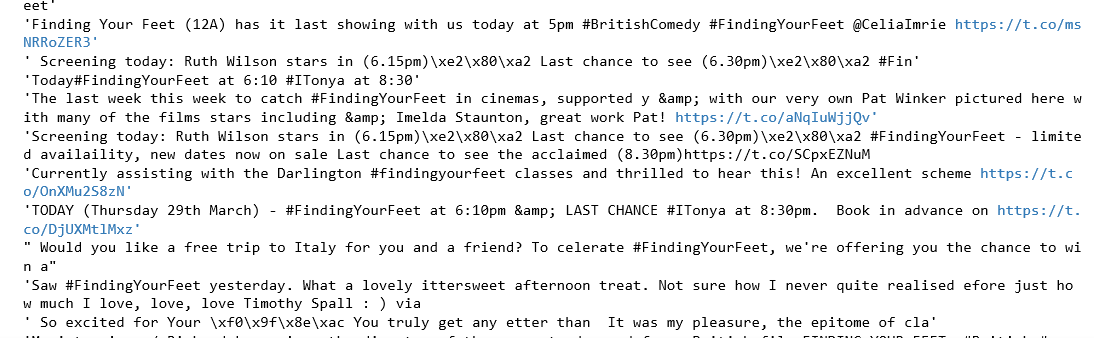
## Data Collection

We have used the Twitter API to crawl the tweets about 14 films on Twitter. And we have collect the data about each film for three weeks on a weekly basis. For each tweet, the tweet’s retweet count, favorite count and post time and the user’s username and number of followers are also crawled.

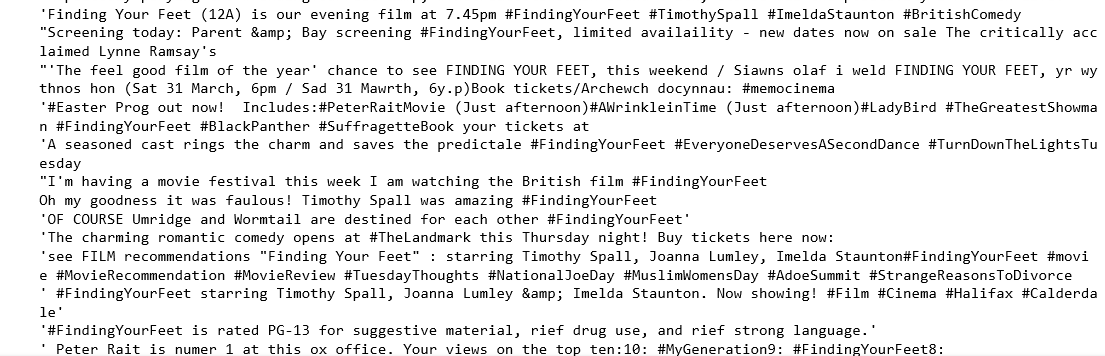
## Data Processing

1. Remove stop-words & Sentiment Score

First, we have removed stop words in the tweet such as icons expression “\\xe2\\x80\\x98s” and the websites containing “http”. Then we have used the Sentiment Analyzer in NLTK package to calculate each tweet’s score and we will use the compound score as a variable in the model.



**Figure 2 - Before removing stop-words**



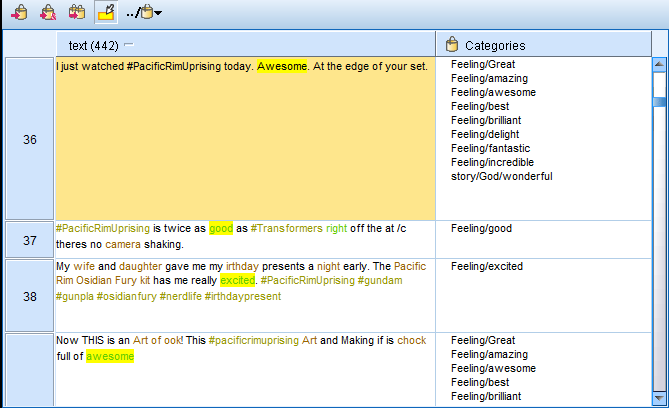
**Figure 3 - After removing stop-words**

1. Tokenization & Tweet Labeling

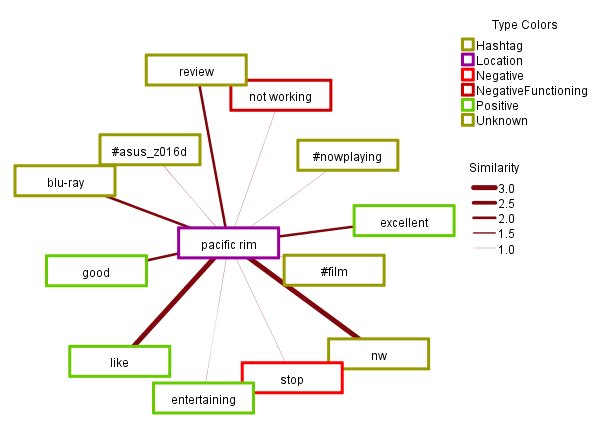
We have tokenized the tweets and calculated the term frequency of the films. After using stop-words in NLTK package to remove stop words, we have picked the first 200 words

We manually created 12 categories related to the concept of movie including: story, director, film type, actor, ticket, theme, promotion, box office, price, feeling, music and platform. Then we used tokenization to split sentences into tokens. By extracting some most frequent words related to our categories from tweets, we labeled them into subcategories.

As can be seen below, the content of tweets about *Pacific Rim: Uprising* is displayed by feelings category. Basically, there are positive feelings such as awesome, good and excellent.



**Figure 4 - Categorization: Pacific Rim Uprising**



**Figure 5 - Map: Pacific Rim Uprising**

1. Process Other variables: release time, box office, budget

Release time is categorized to -1, 1, 2 representing the tweet is posted one week before the release of the movie or within one week or within two weeks.

For Budget and Box Office, we leveraged them into different levels for the prediction model as Table 4 and Table 5 show. The bigger the number of a level, the more money a level represents.

**Table 4 - Levels of Box Office**

|  |  |
| --- | --- |
| **Box Office** | |
| **Standard** | **Level** |
| < $ 50,000 | 1 |
| $ 50k - $ 1million | 2 |
| $ 1 million - $ 10 million | 3 |
| $ 10 million - $50 million | 4 |
| $ 50 million - $ 1 billion | 5 |
| > $ 1 billion | 6 |

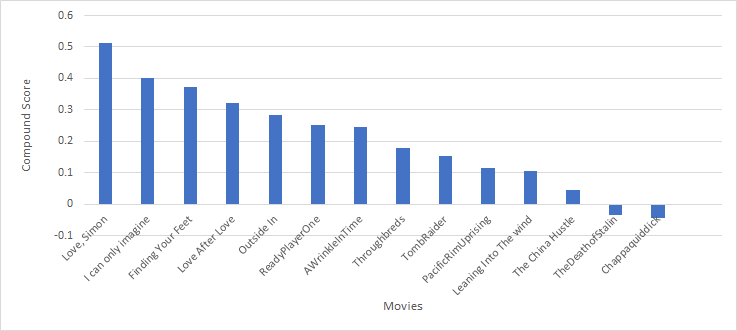
**Table 5 – Budget**

|  |  |
| --- | --- |
| **Budget** | |
| **Standard** | **Level** |
| < 10 million | 1 |
| $10 million - $50 million | 2 |
| $ 50 million -1billion | 3 |
| $ 1 billion - 2 billion | 4 |

1. Merge

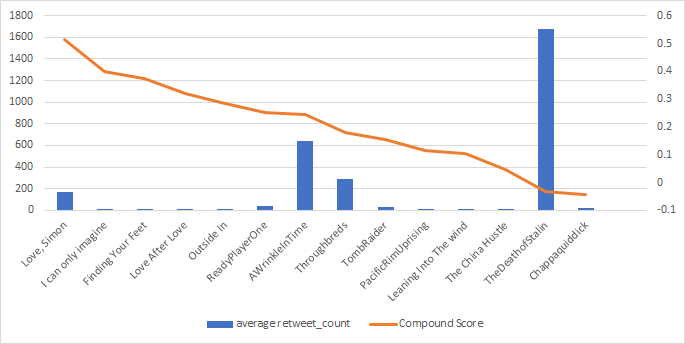
We have merged the cleaned tweet text, tweet information including Retweet Count and Favorite Count and film information using V-Lookup function in excel and we have got our final dataset. Then we have done some descriptive analysis using Tableau.

## Descriptive analytics



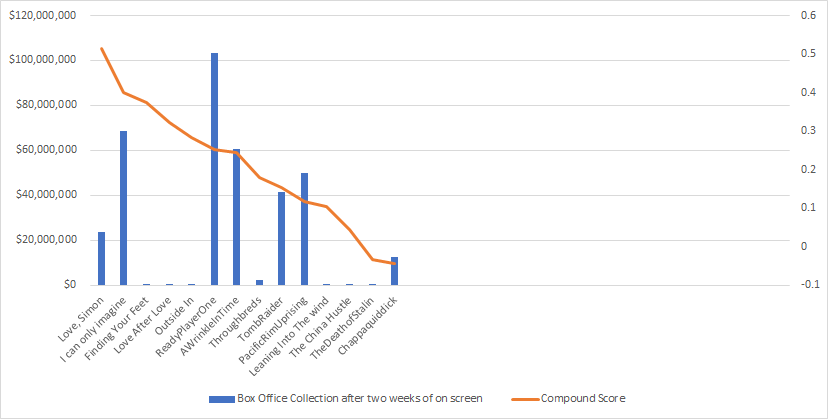
**Figure 6 - compound score across movies**

Most of movies have positive sentiment score. The movie ‘Love, Simon’ has the highest sentiment score but the movie ‘Ready Player One’ has the largest box office collection. There is a difference of about 0.6 between the highest and lowest sentiment scores.



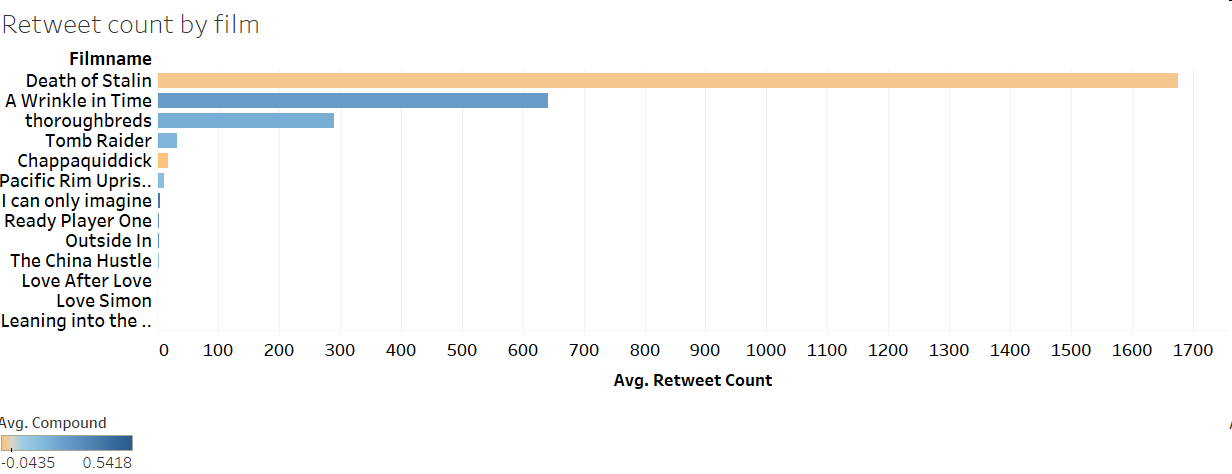
**Figure 7 - retweet count, compound score correlation**

We can see that a movie with lowest sentiment score has a very high retweet count while another movie with around the same sentiment score has a low retweet count, which could be because when a highly anticipated movie gets a bad review, it tends to spread very quickly.



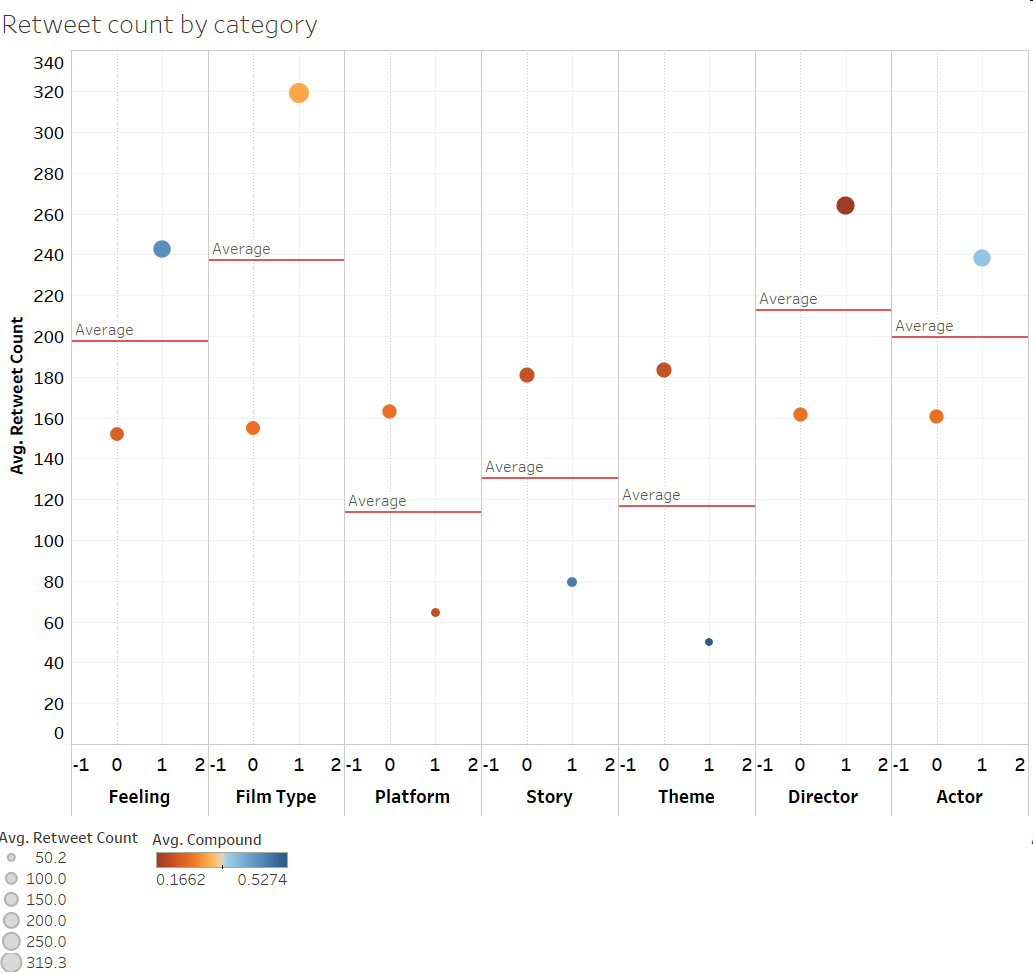
**Figure 8 - box office collection, compound score correlation**

There is no specific relationship between the sentiment score and box office collection.



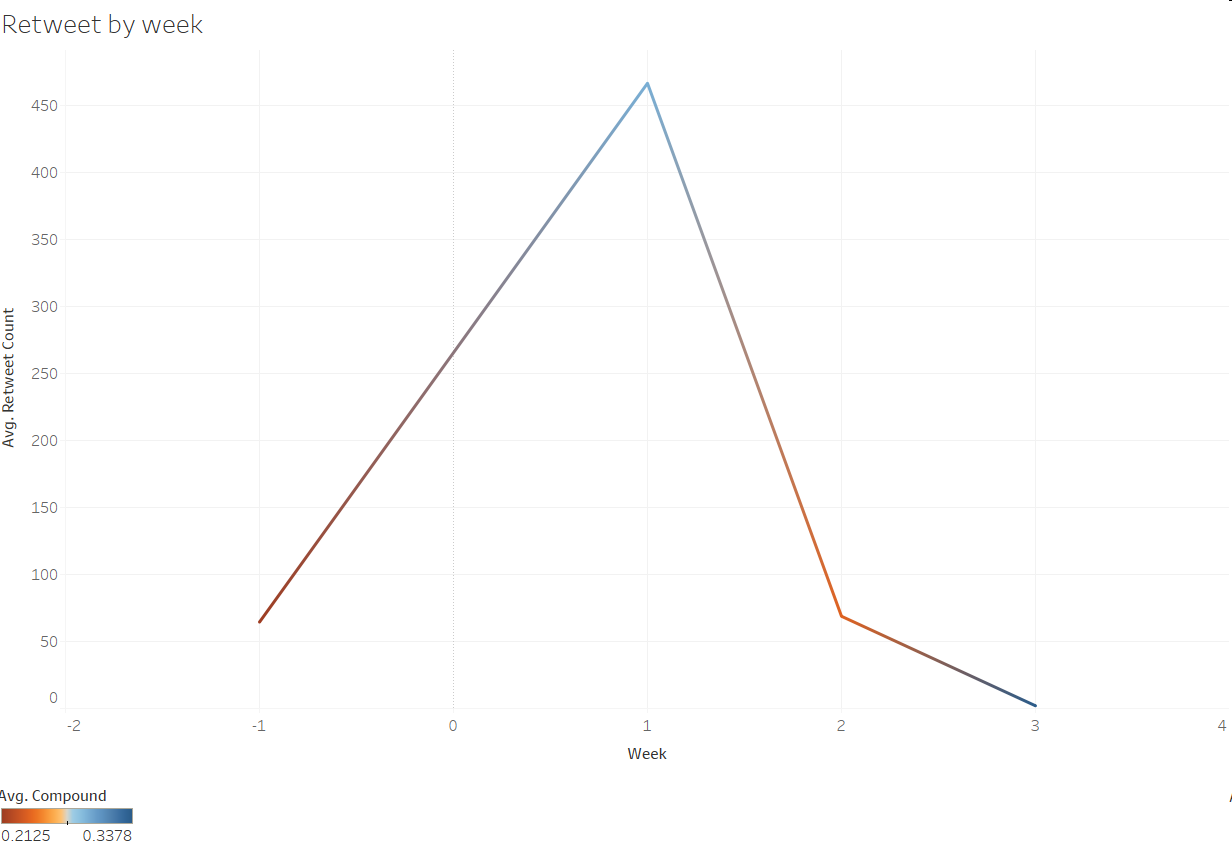
**Figure 9 - Retweet count by film**

The movie DEATH OF STALIN has the highest retweets of all the movies considered. But the compound score of this movie is not proportional to its retweets. We failed to identify a correlation between compound score and retweets.



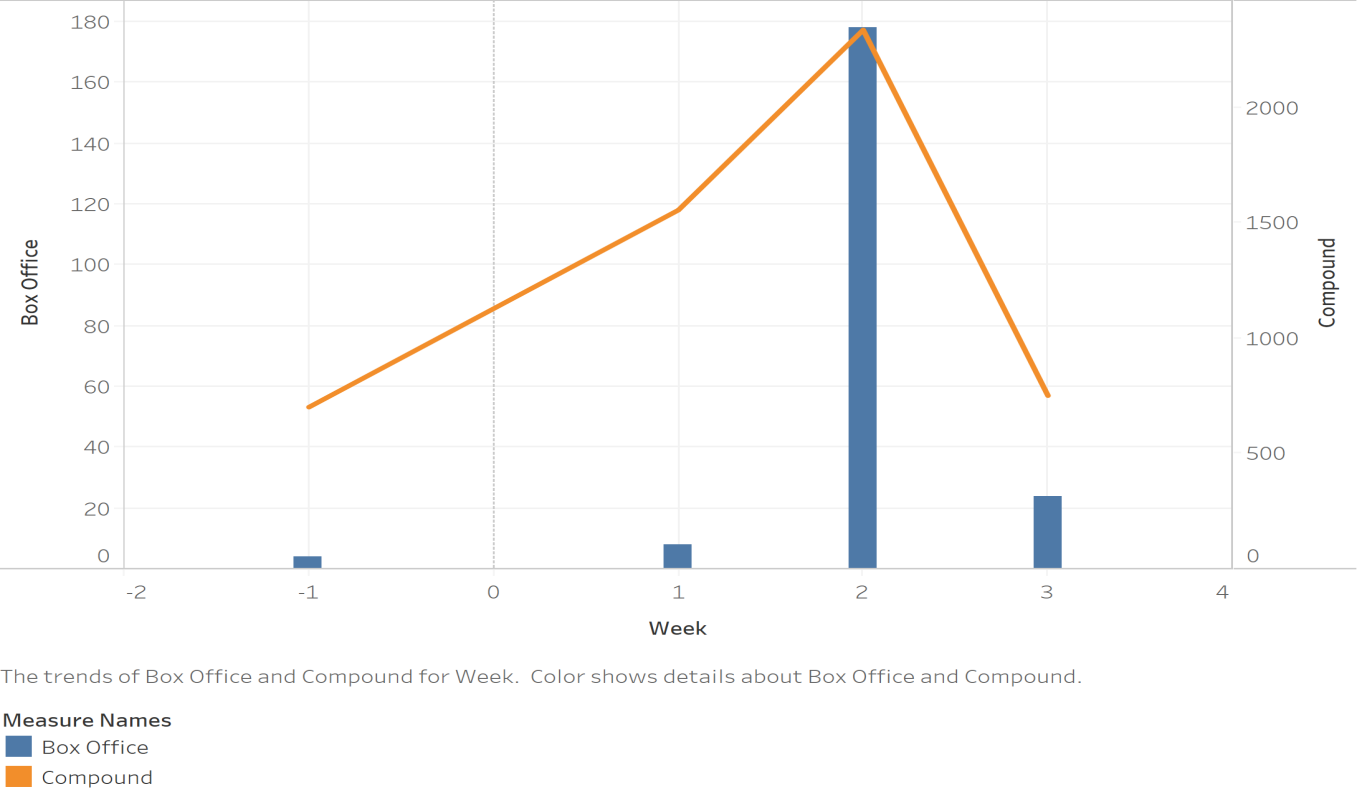
**Figure 10 - Retweet count by category**

Checking the relationship between retweet count and tweet’s category, we can find that tweets talking about film type, director, actor and feeling have higher retweet count. An interesting finding is that the tweets about feeling, film type and actor usually have higher retweet count when it is more positive, while tweets about director have higher retweet when it is more negative. However, this phenomenon varies among different film.



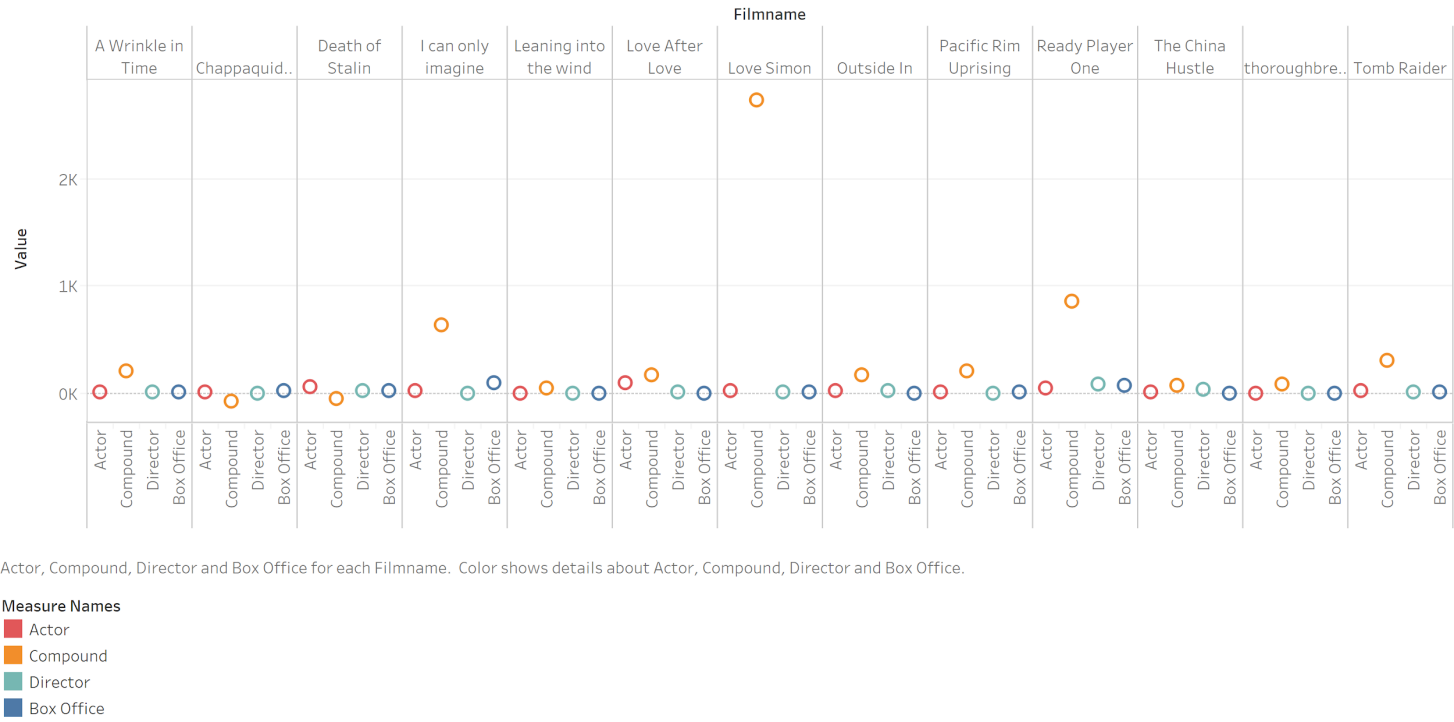
**Figure 11 - Retweet count by week**

Average retweet count also varies in tweets post time. In general, tweets posted in the first week after the film released have higher average retweet count and usually have high sentiment score. In the second week, the retweets dropped down dramatically and goes fewer.



**Figure 12 - compound score, box office collection trend over the weeks**

If we look at the trend of distribution of compound score and box office collection over the weeks before release (-2, -1) and later, there is a steady growth in both the cases. Week 2 seems to be very important as the total compound score and box collection is the largest in this week, even though the distribution of compound score and box office collection vary across movies.



**Figure 13 - actor, director ratings vs compound score, box office collection**

We cannot figure out a correlation between the ratings of actor and directors and sentiment score and box office collection. At Least for the given 14 movies, actor, director ratings do not play a significant role.

## Sentiment Analysis

[Sentiment analysis](https://en.wikipedia.org/wiki/Sentiment_analysis) is simply the process of working out (statistically) whether a piece of text is positive, negative or neutral. The majority of sentiment analysis approaches take one of two forms: polarity-based, where pieces of texts are classified as either positive or negative, or valence-based, where the intensity of the sentiment is considered.

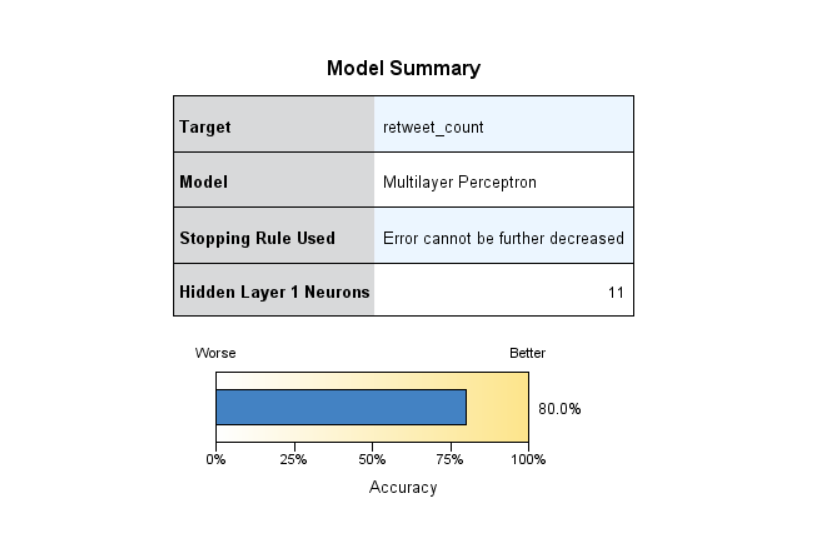
We used Python package VADER Sentiment Analyzer to find the compound sentiment score for each movie(took the average compound score of all the reviews for each movie).Compound score is the sum of all the lexicon ratings ,standardized to range between -1 and 1.We need to load the SentimentIntensityAnalyser object in from the VADER package and used the polarity\_scores() method to get the sentiment metrics for a piece of text. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and works well on texts from other domains.

## System Implementation - Build Prediction Model

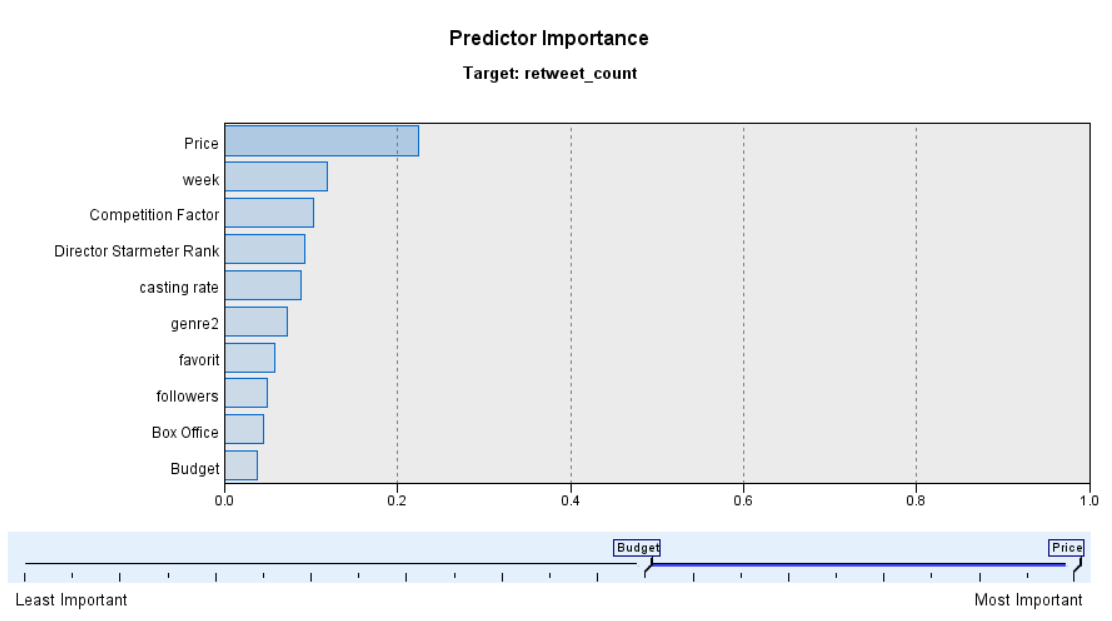
Three models including CHAID, neural network and clustering are used to predict tweets’ retweet count and explore the most important factors to predict a tweets’ retweet count. Neural network has the best performance with accuracy at around 80%.

### Neural network

The neural network has accuracy at nearly 80% and the most important predictors are price, week and competition factor as it shows in figure 8 and 9. In this model, price as a tweet category became the most important factor to predict a tweet’s retweet number, which is a surprising find for us since not many tweets are labeled as price.

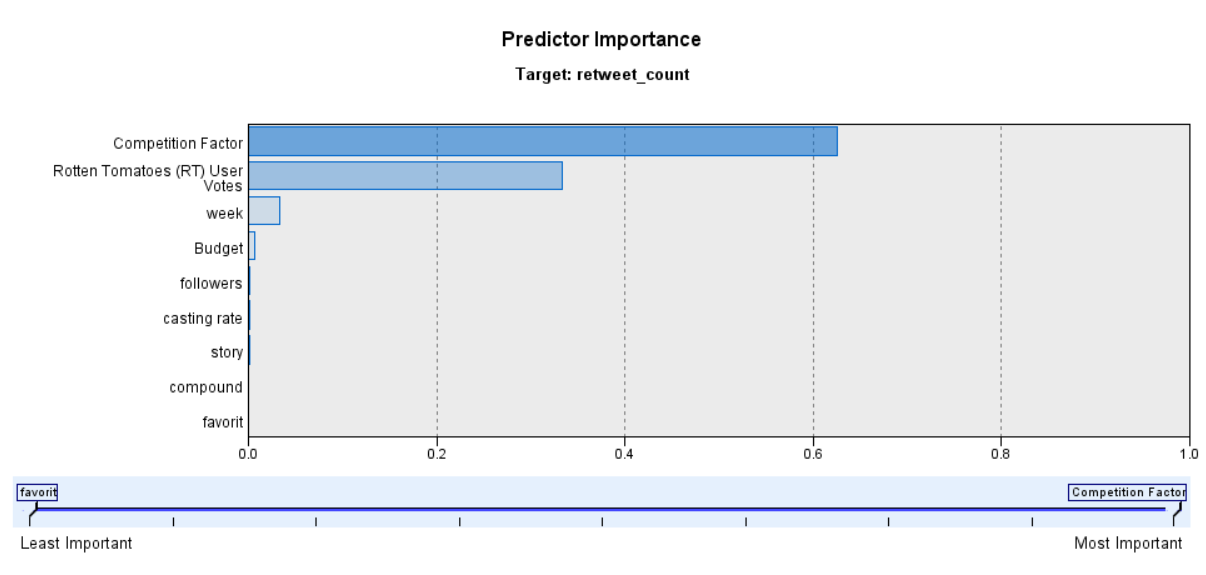
****

**Figure 14 - Neural network model summary**

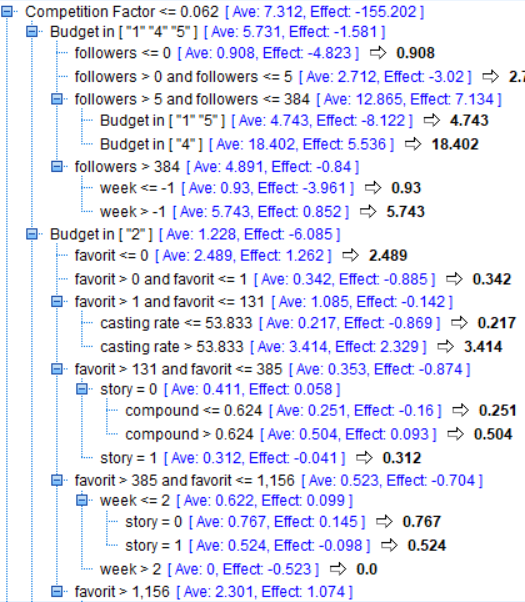
****

**Figure 15 - Neural network predictor importance**

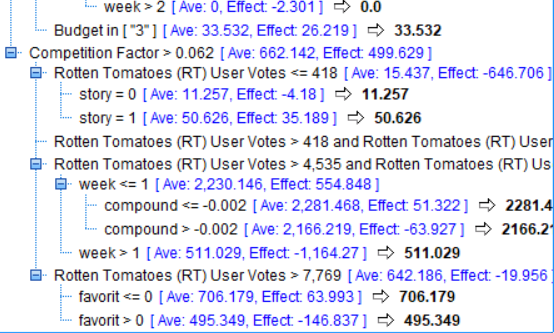
### CHAID

****

**Figure 16 - CHAID model predictor importance**

****

**Figure 17 - CHAID summary-1**

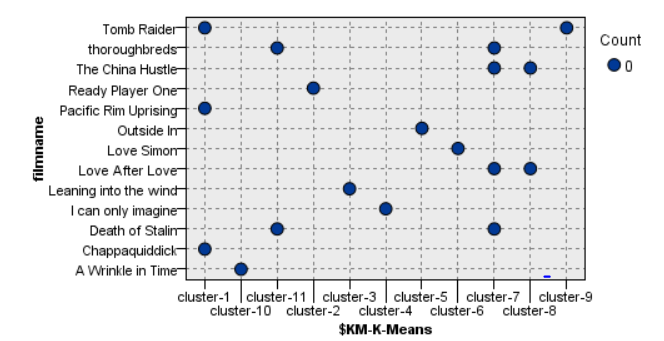


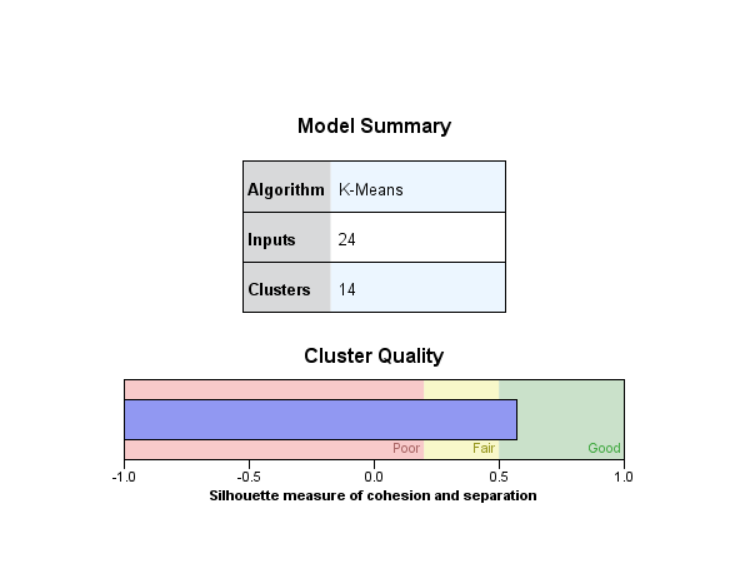
**Figure 18 - CHAID summary-2**

In CHAID model, the most important factors for determining retweets are competition factor and rotten tomatoes user votes. In over case competition factor is almost similar for all the movies. As a result, the outside website rotten tomatoes’ user vote can be detrimental in retweets count which can help in predicting box office collection.

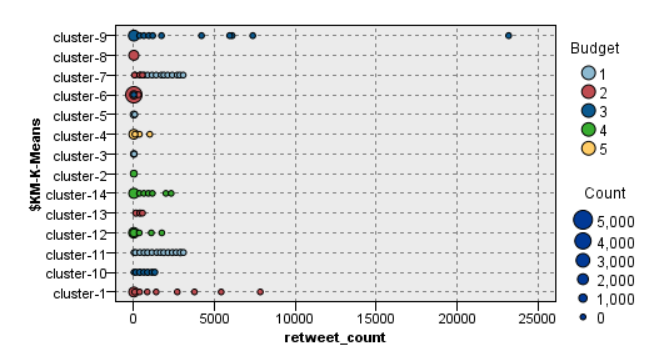
### Clustering

For clustering, we found that when the cluster is over 10, the model performs good and each cluster contains 1 to 4 movies. Figure illustrates the relationship between the clustering result and retweet count, and it is obvious that the clusters cannot used to predict the exact retweet level, but it did discover some patterns of retweet count distribution.





**Figure 19 - Clustering summary & cluster by film**



**Figure 20 - Clustering result and retweet count**

# Conclusions

Initially, this project looked at the impact of text data both before and after film release on volume of retweets. We crawled 20,633 film tweets of fourteen ongoing movies week by week from Twitter. In the data preprocessing, we cleaned and combined the content of each tweets into one dataset. After date preprocessing, customized categories were applied to the cleaned dataset. We analyzed the result using Decision Trees and Neural Network, training the algorithm to predict what was the most important category for volume of retweets. Ultimately, compared to K-means clustering, Neural Network and Decision Tree, Neural Network performed better with nearly 80% accuracy. In result, our analysis, we found that competition factors and weeks are the most important factors to predict retweeted number.

We also found that tweets involving positive words of feeling, film type and actor and tweets involving negative words of director usually brought higher possibility to be retweeted, which indicates that bad news regarding director will be spreading faster.

When looking at the relationship between compound score and box office collection before and after film release date, Week 2 seems to play more important role on box office collection. The findings are expected to provide some suggestions on the time-periodic management of tweets for interested parties wishing to maximize movies’ box office.

# Limitations

Given more time, we would like to optimize our models by enhancing our process to dual with limitations below:

1. The sentiment analysis is depending on text data (7 days before and after movie release) extracted from tweets. The same analysis done on tweets from different time frame can yield different results. Also, different data preprocessing steps can alter the results for same analysis.
2. Most tweets we have collected have low retweet number and the dataset is not normal distributed, so we could not use the linear regression model. In future analysis, we could collect more data more filter the dataset or collect more tweets with higher retweet count and categorize the retweet count to different levels, which will probably make the model more accurate and efficient.
3. The retweeted number is determined by various factors and combination of these factors. The categories in the model can be enhanced by refining categories words to improve the accuracy and quality of the prediction.

# References

1. Zhang, W., and Skiena, S., 2009. “Improving Movie Gross Prediction through News Analysis”. In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 301-304, IEEE Computer Society, Washington, DC, USA
2. Kaur, A. & Nidhi, A.P., 2013. “Predicting movie success using Neural Network”. International Journal of Science and Search, India, online, vol. 2(9), pp. 69-71
3. Delen, D. & Sharda, R., 2006. “Predicting box office success of motion pictures with neural network”. Expert Systems with Applications, Elsevier, vol. 30 (2), pp. 243–254
4. <https://www.analyticsvidhya.com/blog/2017/03/measuring-audience-sentiments-about-movies-using-twitter-and-text-analytics/>
5. <http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html>

# Appendix I

**Table 6 - Tweet Label and Labeling Keywords**

|  |  |
| --- | --- |
| **Label Category** | **Sub-Category Keyword** |
| **Story** | Bisexual (gay, lesbians), story (script, plot, storyline), scandal (implication), unsetting, slapstick, Khrushchev, patriot, starvation, Ukraine, Authoritarian, joseph Stalin tyrant, partys, farce, socialism, tragedy, proletari, died, death, capitalism, absurdist, turmoil, dark, Committee, Soviet, women, Christian, God (faith, testimonies, love, wonderful), Robot (Monstes, kaiju, monstrous), accident, Kennedys (Ted, Kennedy, Teddy), |
| **Director** | Directors, lynn shelton, Lynn Shelton, Speilberg, SteenSpeilberg, Armando Lannucc, Andy Erwin, Ava DuVernay |
| **Film Type** | Romantic-comedy, rom-com, dramedy, Romantic Drama, sci-fi, action, documentary, Comedy-thriller, satire, historical, drama, Action |
| **Actor** | Imelda staunton, timothy spall, actors, cast, celia imrie, zenday, saoirse ronan,  shawn mendes, Lynn Sheltons, performances, role, Justin Timberlake, Jay Duplass Parolee, EdieFalco, funny, jayduplass, Andie MacDowell, Russell Harbaugh, AndieMacDowell, star, TyeSheridan, OliviaCooke, MarkRylance, , BenMendelsohn, EasterEggs, James Halliday, Anya Taylor-Joy, Anton Yelchin, Olivia, jasonsfolly, Steve Buscemi, Jason Isaacs, Rupert, , AliciaVikanders, character, Angelina, casting, Quaid, bDennis, DennisQuaids, Figure |
| **Ticket** | tickets, nTickets |
| **Theme** | love, loss, family, friendship, themes, older, marriage, gay-themed, events, mortality, DnD, game, gaming, TheOasis, scenic, finance, fraud, comedic, dark comedy, historical drama, comedy, FAMILY, Love |
| **Promotion** | 3 years, support, matt bomer, http, recommend, trailer, one week, announced, Trailer, advertising, Download, re#boxoffice |
| **Box Office** | $, box office, commend |
| **Price** | Bill, tax |
| **Feeling** | mood, cutiest, wow, crush, excited, cried, incredible，cry， sweet，omg， shit， feel, amazing, tears, cute, crushed, hopeful, stoked, cry, empathy, obsessed, awesome, incredible, loved, best, Great, cool, impressed, pumped, inspired, quirky, funny, funniest, brilliant, comic, entertaining, delight, hard, dope, exciting, enjoyed, liked, better, pleasure, want, AWESOME, worth, fantastic, bWoo, privilege, movie of the year, bGreat, well, Thanks, good, unexpected, worthwhile, mind-blowing, Glad, fantastic, thoughtful, frustrated, disappointed |
| **Music** | Music, Erik, Friedlander |
| **Platform** | Twitter, Netflix, netflix, via, AMCTheatres, itunes, PS4, Switch, Warner Bros, Xbox, iTunes, Hulu, HBO |

# Appendix II

**Figure 12 - Clustering Information**

