National College of Ireland

# Early prediction of a film's box office success using natural language processing techniques and machine learning

MSc Research Project
Data Analytics

Sean O'Driscoll
x15001288

School of Computing
National College of Ireland

Supervisor: Dr. Dominic Carr

**National College of Ireland**
**Project Submission Sheet – 2015/2016**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Sean O'Driscoll |
| **Student ID:** | x15001288 |
| **Programme:** | Data Analytics |
| **Year:** | 2016 |
| **Module:** | MSc Research Project |
| **Lecturer:** | Dr. Dominic Carr |
| **Submission Due Date:** | 12/12/2016 |
| **Project Title:** | Early prediction of a film's box office success using natural language processing techniques and machine learning |
| **Word Count:** | 6769 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 12th December 2016 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Early prediction of a film's box office success using natural language processing techniques and machine learning

Sean O'Driscoll
x15001288
MSc Research Project in Data Analytics

12th December 2016

## Acknowledgements

I would like to thank my supervisor Dr. Dominic Carr. His help and advice with all aspects of this research project was an invaluable resource.

## 1 Abstract

This research applied natural language processing and machine learning techniques to film scripts in order to try to predict whether or not the film will be financially successful. The film scripts were transformed into a term document matrix, with term frequency-inverse document frequency scores used to assign feature importance. The machine learning algorithms used in this research were decision trees, random forest, naive Bayes, and support vector machines. The results were evaluated using accuracy, precision, recall, F1 score and where appropriate, Cohen's Kappa. The results were also compared to predictions made using information about the films that is either known or can be reasonably estimated before the film has been made. Film scripts were also analysed after first segregating them by genre, in order to compare scripts with more similar/ related material. Overall, the predictions made using data generated from the film scripts were poor, while the predictions made using information about the films were only slightly better, based on this research's stringent evaluation criteria.

# Contents

# 2 Introduction

An incredible amount of money is spend on entertainment, ($320 billion dollars annually by Americans) (Vogel; 2014). According to PwC, the revenue generated by the film industry in the US in 2014 was $32 billion dollars (Statista; 2016a). The large sums of money involved make it understandable that tools that aid in the prediction of box office results would be of great value to potential investors. It is very difficult to accurately predict if a film is going to be a financial success before it has been made. Academy award winning screenplay writer William Goldman claimed that nobody knows with certainty if a film is going to be a success or not and at best people make educated guesses (Goldman; 2012).This research focuses on using screenplays to predict whether or not a film will be financially successful as they represent the first and arguably most important step in the film making process. The screenplays are essentially the blueprint upon which the film is based (Nelmes; 2007).
There are a number of research gaps or weaknesses in the related research that this study attempts to fill. These include;

- The lack of consideration of international box office results
- The very small sample sizes used in related studies
- The lack of strict adherence to only using information that is available or can be reasonably estimated before a film has been made
- This research is the first to analyse film scripts after first segregating them by genre

The remainder of this paper is organised as follows; Section 3 consists of an up to date literature review which briefly assess the relevant research in this area in order to provide context for this research. Section 4 outlines the methodology used when completing this research. With the information provided in this section and the user configuration manual, it should be possible for others to recreate/ repeat this research. The results of the various models are evaluated in Section 5. This is then followed by a conclusion which summarizes the findings of this research and provides ideas for potential areas of further work in Section 6. Appendix A contains various tables of results from the different models and data used.

# 3 Literature Review

One way to categorize the many different approaches to predicting box office results is to examine the time line of predictions.The various different prediction methods fit into one of three categories;

- Early predictions; predictions which are made before the film has been made
- Hype-based predictions; predictions made after the film has been made but before it has been released
- Late predictions; predictions made after the film has had its initial release

Table 1.0 summarizes which category each of the papers reviewed belongs to.

Table 1: Prediction timetable

| **Early predictions** | **Hype-based predictions** | **Late predictions** |
|---|---|---|
| Burgos et al. (2005) | Zufryden (2000) | Jedidi et al. (1998) |
| Eliashberg et al. (2007) | Krauss et al. (2008) | Neelamegham et al (1999) |
| Goetzmann et al. (2013) | Zhang and Skiena (2009) | Simonoff and Sparrow (2000) |
| Eliashberg et al. (2014) | Asur and Huberman (2010) | Elberse and Eliashberg (2003) |
| Ghiassi et al. (2015) | Mestyán et al. (2013) | Sharda and Delen (2006) |
| Hunter et al. (2016) | Kim et al. (2015) | Lee and Chang (2009) |

## 3.1 Early predictions

The price paid for the screenplay has been shown to predict the success of the film (Goetzmann et al.; 2013). Only 151 of the screenplays used in this study were produced into films. This represents a small sample size. By analyzing the scripts of films using natural language processing techniques, Eliashberg et al. (2014) were able to predict the box office results of films. This research built upon (Eliashberg et al.; 2007) work which used 1 page summaries or spoilers instead of scripts. One shortcoming of the method used by (Eliashberg et al.; 2014), is that it requires the scripts to be read by two people who have expert domain knowledge of screenplays. Although (Eliashberg et al.; 2014) claim their prediction method is just another decision aid that studio heads could use when deciding which films to green light, surely one of the goals of developing such tools should be to reduce the reliance on humans with expert domain knowledge? The goal of this research is to make box office predictions based on scripts without the input of domain experts whom have read the script. By using network text analysis, i.e. representing the films scripts as a network of interconnected concepts (Hunter et al.; 2016) showed that the size of the text network is positively associated with box office performance. A possible drawback for using the size of the text network of a script as the basis for green-lighting a film is the difficulty in explaining to a producer, who is not familiar with network text analysis, why a script with a larger text network is better than a script with a smaller text network. Also, once writers know that larger text networks are preferable than smaller text networks, they can game the system by purposefully including an increased number of multi-morphemic compounds and trying to reverse engineer a larger text network for their script. It is possible to produce an extremely accurate box office prediction tool which relies solely on information available during the pre-production phase, such as Ghiassi et al. (2015)'s model which has an accuracy of over 90%. However, as Ghiassi et al. (2015)'s model is a dynamic artificial neural network, it is a black box method and therefore its reasonings are not easily understood or explained. This means that, like the network text analysis method, it would be difficult to use this method to convince a film producer of the box office potential of a script. The predictions made by Ghiassi et al. (2015)'s do not take return on investment into account, only box office takings.

## 3.2 Hype based predictions

By predicting the box office performance of movies using news data from online daily newspapers, (Zhang and Skiena; 2009), were able to predict box office results as accurately as others whose methods used categorical data about the film from the website *IMDb* (2016). As the predictions were based on news articles from before the film was released, this method of prediction provides earlier results than methods that are based on opening

weekend box office results. It is clear from the findings of (Zhang and Skiena; 2009), that the prediction of a movie's box office results from analyses of news articles, (both sentiment analyses and article counts), works best on high budget movies that are likely to be released on a large number of theaters and have high box office grosses. Similar box office prediction methods include those that analyze Twitter data related to a film, (Asur and Huberman; 2010), measure the amount of views and edits of the films Wikipedia article, (Mestyán et al.; 2013), analyzes of movie form discussion, (Krauss et al.; 2008), or use the amount of activity on the films website, (Zufryden; 2000). The obvious drawback of these methods is that they are only useful for predicting the box office results of films that have already been made and therefore, not useful for deciding whether to green-light production of a film or not. In some cases, such as if the film is an adaptation of a popular book, or a sequel to a previous film, analyzing the amount of excitement surrounding the film, (via analyses of tweets, Wikipedia page activity, the films own website activity or through news articles), could possibly be used to predict the box office results before the film has been produced. This would then be considered a method of early box office prediction.

## 3.3 Late predictions

A variety of machine learning techniques have been employed in attempts to try to predict box office performance. These include clustering, (Jedidi et al.; 1998), neural network, (Sharda and Delen; 2006), Bayesian Belief networks, (Neelamegham and Chintagunta; 1999), (Lee and Chang; 2009) and regression modeling, (Simonoff and Sparrow; 2000). In the case of these studies, information that would not be available until after the film is released, such as critics' reviews (Elberse and Eliashberg; 2003) or award nominations, is used in the prediction models. As approximately 25% of total box office gross is generated during the film's opening week (Simonoff and Sparrow; 2000), a weakness of research that focuses on predictions after initial release is that a large fraction of the box office gross has already been generated at this point. Another weakness of this area of research is that the predictions come far too late to influence the decision of whether or not to finance the film.

## 3.4 What exactly are these studies trying to predict?

By converting the box office results prediction problem into a categorization problem, (Sharda and Delen; 2006), drastically reduced the number of potential results the predictions could have. By having 9 categories, which range from flop' to 'blockbuster', Sharda and Delen (2006)'s prediction model can give an adequate indication to film executives as to how the film will perform, without having to give a point estimate of the actual box office result, which ranges from close to $0 all the way up $2.8 billion for Avatar (www.the-numbers.com, 2015) in 2009. Others who use a similar classification system include (Zhang and Skiena; 2009), although their system has only 6 categories. The drawback of these methods is that films with large production and advertising budgets are likely to have large box office grosses. However, this does not mean that these films are profitable. By only categorizing films as profitable or not profitable, (Burgos et al.; 2005), are able to report an accuracy of 72.66% from their model, which uses decision trees to predict which category a film will fall into. 87.5% accuracy was achieved by (Simonoff and Sparrow; 2000). However, their results are based on only having predicted 21

out of 24 movies correctly, (a very small sample), while they also use an extremely large prediction interval. The prediction interval is over $150 million wide in one instance. The drawback of using profitable or not profitable is that it only tells us that the film either made more than, or less than the costs associated with making the film. A prediction model which makes predictions based on potential return on investment represents a more useful tool. This research aims to categorize films as either successful or not successful based on meeting a ROI ratio above 1:1.
In order to provide a comparison and context to the prediction method outlined above, two alternative methods will also be used. These alternative methods will consist of categorization methods similar to those used by Sharda and Delen (2006), but will use 3 and 5 categories respectively. Cohen's Kappa (Cohen; 1960) will be used to evaluate these categorisation approaches.

# 4 Methodology and Implementation

The methodology used in this research consisted of the following steps:

- Data collection
- Data transformation
- Application of models
- Evaluation of results

## 4.1 Data Collection

The data used for this research consisted of film scripts collected from the websites *The Internet Movie Script Database* (2016), *The Daily Script* (2016), *Screenplays For You* (2016), and data about the films collected from *The Numbers* (2016). The scripts were scraped from the websites using the BeautifulSoup package in Python. The box office results, budget, release date, genre and MPAA rating had to be recorded manually.

## 4.2 Data Transformations

The film scripts had to be cleaned and transformed into a format that was suitable for applying machine learning techniques. This included the removal of any non-letter characters in the script. This was done using regular expressions in Python. In the cases where the film scripts contained additional words, due to the content of advertisements being unintentionally scraped, Notepad++ was used to find and delete these superfluous sections of the scripts. Once the film scripts had been transformed into their cleaned forms, they were tokenized, stemmed and converted to lowercase (Rahm and Do; 2000). This was done using the NLTK package in Python. The reason for stemming words and converting them to lower case is so that the importance of a word is not diluted by having it in several different forms (Paice; 1994). For example, the words 'running', 'runner', 'Run', 'run' and 'runs' will all be represented by the word 'run' after stemming and converting to lowercase has been completed.

Next, the term frequency-inverse document frequency, (TF-IDF), was calculated for all

of the scripts. This was completed using the TfidfVectorizer from the scikit-learn package in Python. The maximum number of features to be considered by the TfidfVectorizer was set at 1000. A possible area of further study could include changing the number of features considered. The 1000 features considered by the TfidfVectorizer were printed out so that they could represent column headers in the TF-IDF term document matrix once that was created. The TF-IDF scores were outputted into a single column of numbers, the first 1000 of which representing the scores for the first film script, the second 1000 representing the scores for the second script and so on. This column of numbers was converted into a 922 x 1000 term document matrix. This was done in Excel using the following formula;

$$= INDEX(\$A : \$A, ROW(A1) * n - n + COLUMN(A1)) \tag{1}$$

Where $n$ is the number of features in the term document matrix.

TF-IDF has been shown to be successful at determining word relevance in the area of document queries. Its advantages include that it is efficient, simple and easy to implement (Ramos; 2003). The disadvantages of TF-IDF include that, as a result of considering each word independently, in fails to see the association between words that are synonyms. TF-IDF would also consider words such as worker, working, works and worked as different, independent words. This limitation can be partly overcome by stemming the words. However, the failure to see that words such as clever, smart, astute and intelligent could be considered interchangeable and therefore should not be counted independently cannot be easily remedied. One way in which this limitation could manifest itself in this research is if film scripts for various films set during wars mentioned tanks, but each film referred to the specific model tank model such as Sherman, M67 Patton or the M1 Abram. These scripts may have a very strong similarity that the TF-IDF score would not detect (Ramos; 2003).One potential method for dealing with this issue is to consult special synonym dictionaries (Rahm and Do; 2000).

While a word matrix was created using the TF-IDF scores calculated using all of the film scripts combined, additional versions were created using only film scripts that belonged to the same genre as each other. This represents a new way of using natural language processing to analysis film scripts. The reasons for doing this are discussed further in section 4.4 Separating Scripts by Genre.
The budget data for each film was converted to the 2014 equivalent using the consumer price index. It was important to do this because, a film from 1974 with a budget of $13 million dollars would be encoded as a 1 or very low budget film unless inflation is considered. After taking inflation into account, the budget would be over $62 million dollars and thus is encoded as a 3 or medium budget film. Film budgets where adjusted for inflation using the following formula;

$$x = (b) * (CPI2014/CPIy) \tag{2}$$

where;
$x$ is the film budget adjusted for inflation.
$b$ is the budget the year the film was released.
$CPI2014$ is the consumer price index for 2014

$CPIy$ is the consumer price index of the year the film was released
(Appelbaum; 2004)

Each film used in this research was categorised as either 'successful' or 'not successful' based on the film's return on investment. If the film's return on investment, calculated using the film's unadjusted budget and unadjusted domestic box office return, was greater than or equal to 1, the film was categorised as 'successful'. If the film's ROI was less than 1, the film was categorised as 'unsuccessful'. If more data was available, films could be categorised into a larger range of categories such as 'moderately successful' and 'very successful'. This represents a possible area for future research.

### 4.2.1 Latent Semantic Analysis

Another strategy for predicting box office success involved reducing the data in the term-document matrix. This was done using latent semantic analysis (LSA). LSA is based on singular value decomposition. Like the bag-of-words model, LSA does not take word order into account (Landauer et al.; 1998).

Initially, when the 1000 feature term-document matrix was reduced to 100 concepts, 98 of the concepts only contained names. This revealed the need add names to the list of stop-words that were removed from the scripts. A list of approximately 30,000 first names was added to the list of stop-words that were removed from the scripts. The term-document matrix was recreated and then reduced to 100 concepts. Latent semantic analysis was used by (Eliashberg et al.; 2014). The use of only 100 words in the document-term matrix appears extremely small. To represent the 300 film scripts used by (Eliashberg et al.; 2014) in their study with only 100 words, which was then reduced to 2 dimensions using latent semantic analysis, seems like an area that is worth exploring further. This research represents 922 films scripts with a term-document matrix that contains 1000 features. These 1000 features were then reduced to 10 dimensions using latent semantic analysis. The explained variance for each of the 10 new concepts was plotted and can be seen in Figure 1.0. There was a clear elbow in the line at third component. For this reason, only the first three components were used for the predictions, as the diminishing amount of variance explained by the remaining concepts did not warrant inclusion. By plotting the first, second and third components, created by the LSA procedure, for both the 'successful' and 'unsuccessful' films, it is clear that there is no correlation between the location of each data point and the data point's label. This can be seen in Figure 2.

Table 2: TF-IDF term document matrix subsection

| abl | abov | across | act | action | actual | address |
|---|---|---|---|---|---|---|
| 0.010 | 0.004 | 0.000 | 0.004 | 0.000 | 0.000 | 0.000 |
| 0.048 | 0.002 | 0.002 | 0.000 | 0.001 | 0.005 | 0.012 |
| 0.019 | 0.007 | 0.024 | 0.065 | 0.053 | 0.025 | 0.023 |
| 0.002 | 0.008 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |

## 4.3 Application of Models

In order to make predictions about whether or not a film will be successful by using the TF-IDF word matrix, an additional column with the target label, 'success', must be ap-

Figure 1: Scree Plot, Latent Semantic Analysis

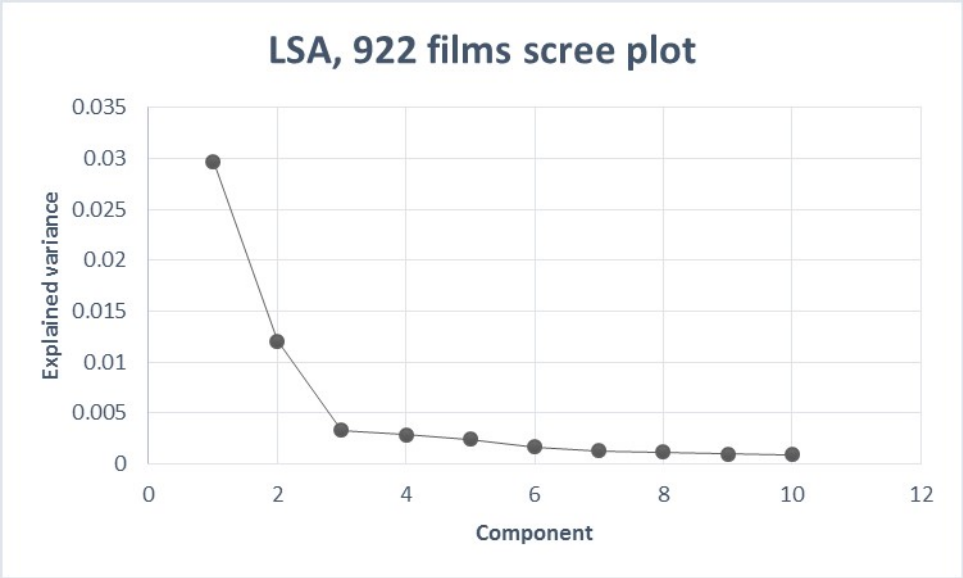
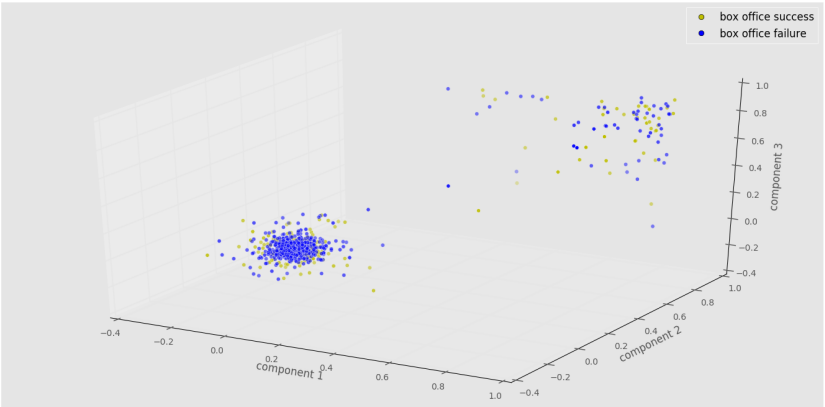



Figure 2: First, second and third principle components, LSA




pended to the matrix. A 10 fold cross validation method was used when applying models. The 10 fold cross validation method has a number of advantages over the alternative hold out method. As there is not an abundance of data available, the 10 fold cross validation method makes best use of the data available, by putting all of the data to use in both the training and testing phases. The 10 fold cross validation method reduces the chances of overfitting the data (Kohavi et al.; 1995).

## 4.4 Separating Scripts by Genre

One approach used to try and improve the results of the models was to separate the films by genre first. Of the research that tries to predict box office results by analysing the script, (Eliashberg et al.; 2014) and (Hunter et al.; 2016), no other research attempts to analysis the scripts after first segregating them by genre. By doing this, comedy film scripts would only be compared to other comedy film scripts while action film scripts

would only be compared to other action film scripts etc. It was hypothesized that the features of action films could be different to that of comedy films and thus could benefit from being investigated separately. The limited amount of data available meant that the only genres considered independently were, 'Action', 'Comedy', and 'Drama'.
A limitation of this approach is that the assigning of a genre to a film can be subjective. For this research the genres allocated to each film by the website *The Numbers* (2016) were used. This was done to maintain consistency throughout the research. However, it should be acknowledged that many films could be assigned to more than one genre.

## 4.5 International Box office results

Related research has primarily focused on the box office results from the United States of America and Canada alone. One of the aims of this research was to take the international box office results into account when making predictions, as this is an increasingly large segment of the movie industry market. The data for international box office results is not as readily available and as such, only 663 of the film scripts could be used when making predictions based in international box office results. While the North American market remains the biggest in terms of box office revenue, ($11 billion in 2015), other international markets such as China are continuously growing and are becoming increasingly important (Statista; 2016b).

## 4.6 Release Date

There is no consensus throughout the research as to how the release date of the film should be considered in the prediction models. The website *Box office Mojo* (2016) uses five seasons to categories release dates. Release month was considered by Burgos et al. (2005), three seasons were used by Simonoff and Sparrow (2000) while a seasonality coefficient, derived from approximately 2000 films released between 1985 and 2000 was used by Ghiassi et al. (2015). Traditionally January-February and August-September have been considered 'Dump Months', where films with lesser expectations have been released (Burr; 2013). However, recent results such as, Guardians of the Galaxy, 2014 and Suicide Squad, 2015 opening at $94,320,833 and $133,682,248 respectively contradict this convention. Also, Deadpool, 2016 had an opening of $132,434,639 in February while American Sniper 2015 opened with $89,269,066 in January (*Box office Mojo*; 2016). These examples show how the film market is continuing to evolve with time. Due to these inconsistencies, the various models were also tested with the release month variable removed.

## 4.7 Data Reduction

In order to try and increase the accuracy of the prediction models, the data used was reduced based on the following criteria:

- Movies release before 1990 or after 2012 were removed.
- Movies with a production budget under 1 millions dollars, (after adjusting for inflation), were removed.

Although one of the aims of this research was to use a far larger sample of data than was typically used in the related research, removing the films from before 1990 and after 2012

reduced the time span of the film's release dates from 52 years to 22 years. It did this while manging to maintain over 80% of the films used in the research. This research assumes that films made more recently are more relevant than films made long ago to films that will be potentially made in the future. Ideally, this research would be conducted using the 100 biggest films, (based on ROI), of each year for the last 10 years. However, the scripts for the majority of these films were not available.
Removing the films with production budgets of under 1 million dollars served two purposes. Firstly, it helped to increase the chances that every film being used in the research reached a minimum production value. (There is no point analysing the script of a film if the final product was made using substandard production equipment and techniques). Secondly, it removed films that where extreme outliers based on their ROI, due to their exceptionally low production costs.

## 4.8 Classification Algorithms

The machine learning algorithms used in this research were decision trees, random forest, naive Bayes and support vector machines. As each of the films used in this research has been labelled as either 'successful' or 'not successful', it is appropriate to use supervised machine learning techniques.This section will briefly explain the algorithms used.

**Decision trees**
Decision trees are a supervised learning method. They mirror human decision making processes. One of their advantages is that they can be explained to non-technically literate people who are involved business decision making process. A large number of features in the data can result in the decision tree overfitting. For this reason, the decision tree algorithm could be more useful for this research once the number of features in the data has been reduced using latent semantic analysis or when using the movie data, (genre, MPAA rating etc.) (Lantz; 2015).

**Random Forest**
Random forest is an example of an ensemble method of prediction, i.e. a method of prediction that uses multiple classifiers and averages the results in order to make predictions (Liaw and Wiener; 2002).Random forest uses decision trees as the classifiers, each of which votes on which category each sample in the test data belongs to (Breiman; 2001).

**Naive Bayes**
Naive Bayes classifier are classifiers which use Bayes Theorem and are built on the bases that the probability of an event occurring in the test data can be estimated based on what data is present or absent in the training data (Lantz; 2015) (John and Langley; 1995). For this research the event occurring will be the film being successful and the data will considered to estimate the probability will be the words from the film's script. Naive Bayes classifiers assume all features are independent which in text classification, they are not (Rish; 2001). However, they considers all of the features in the dataset, not just the features it considers to be the most important. This provides a suitable contrast to the method used by decision trees. Multinomial naive Bayes takes account of word frequencies and has been successfully used in text classification problems Witten and Frank (2005).

**Bayes' theorem**

$$P(H_n/E) = (P(H_n)P(E/H_n))/(\sum_m P(H_m)P(E/H_m)) \quad (3)$$

Where $E$ is any event, $H_n$ is a sequence of exclusive and exhaustive events (Lee; 2012)

**Support Vector Machines**
A support vector machine, SVM, uses a hyperplane or linear decision surface to create a boundary between data points, thus dividing the data points into two separate groups. The SVM tries to find the Maximum Margin Hyperplane, MMH, in order to create the largest possible division between the two categories of data (Lantz; 2015) (Cortes and Vapnik; 1995). SVMs are very well suited to binary classification problems and have been successfully used for the purpose of text classification (Joachims; 1998).

# 5 Evaluation

## 5.1 Accuracy, precision and recall

The results generated from the various models were evaluated a number of ways. The accuracy, precision, recall and F1 score was calculated for each of the models used. These results were compared to the results achieved by simply classifying every film as 'successful'. By comparing the results obtained by the various models to the results obtained by chance, it provides a context within which to consider the results. The computational time required for each model was also considered. Finally, results were compared to the results of applying the same models to additional data that is available or could be reasonably estimated about each of the films used, (movie data). This additional data contains information on the genre, the release month, the Motion Picture Association of America rating, and the budget of the film.

True Positive, (TP) = films correctly classified as 'successful'
False Positive, (FP) = films incorrectly classified as 'successful'
True Negative, (TN) = films correctly classified as 'not successful'
False Negative, (FN) = films incorrectly classified as 'not successful'

$$Accuracy = (TP + TN)/((TP + FP + TN + FN)) \quad (4)$$

$$Precision = TP/((TP + FP)) \quad (5)$$

$$Recall = TP/((TP + FN)) \quad (6)$$

$$F1Score = 2 * ((P * R)/(P + R)) \quad (7)$$

Precision is the ratio of correctly classified positives to the total number of instances classified as positive.

Recall measures, for all of the films that should have been classified as 'successful', how many were actually classified as 'successful'.
Accuracy is the ratio of correct predictions to the total number of predictions made.

When using a binary classifier, if 80% of the test cases are 'positive', and the classifier simply labels all of the test cases as 'positive', the classifier would have an accuracy of 80%. This type of classifier would not represent a useful prediction tool. This is why it is important to consider the precision, recall and F1 score achieved by the classifier. In the case of movie box office prediction, is it more important to have very high recall, i.e. avoid missing a film that will be successful, or have very high precision, i.e. ensure the vast majority of the films predicted as successful will be successful? As the average cost involved with producing a major studio film are extremely high, ($65million dollars in 2007 (Mueller; 2011)), the number of films produced by a studio are relatively low, (only 708 films released in US in 2015 (MPAA; 2015)), and there is a constant supply of potential scripts available to a film studio, this researcher believes it is more important that the classifier has high precision rather than high recall.

## 5.2 Cohen's Kappa

In order to evaluate the results from the classification methods with 3 and 5 possible outcomes respectively, Cohen's Kappa (Cohen; 1960) was used. This evaluation method takes account of the fact that chance agreement occurs. It adjusts the observed proportional agreement which would be expected to occur by chance.

$$k = p - p_c/1 - p_c \tag{8}$$

where;
$k$ = Cohen's Kappa
$p$ = the proportion of units agreement
$p_c$ = the proportion of units which would be expected to concur by chance.
(Cohen; 1960)
It is clear from looking at the results in tables 3-24 in Appendix A that the overall results of the various prediction models on the various different versions of the data are profoundly poor. By first considering the F1 score, which incorporates both the precision and the recall scores of the model, only models applied to the data regarding the films for which international box office results were available, achieved good results. It is not obvious as to why the models performed so much better on this subset of the data. Tables outlining all of the results achieved using a variety of data and prediction algorithms are included at the end of this report in Appendix A. The overall findings from these results can be summarised as follows:

- Movie data provided better results than data generated from analysing the scripts, (script data), when using decision trees, random forest and support vector machines.
- The naive Bayes algorithm produced better results using the script data, although results were still poor.
- Segregating the films by genre before generating scripts data/ applying models resulted in slightly improved results, (but still quite poor).

- Reducing the script data using Latent Semantic Analysis did not improve the prediction results.
- Removing films with budgets under 1 million dollars or release dates prior to 1990 did not improve the results.
- Only considering the 663 films for which the international box office results were known resulted in significantly improved results. By only considering films for which the international box office results were known, any films that didn't get released internationally was automatically excluded. This could have resulted in an increase in the overall quality of the films being considered at this stage in the research.
- Removing the release month variable did not significantly change the results, positively or negatively. Possible reasons for this are discussed in section 4.6 Release Date.

### 5.3 Computation time

Another area considered for the evaluation of the various models was the computation time required. A sample of the computation time required by the various models, using both the script data and the movie data, can be seen in Table 25. From the results in Table 25 it is clear that the script data consistently requires more computation time than the movie data and that the random forest model requires considerably more computation time than the other models tested. This is not surprising, as random forest is an assemble method which constructs numerous decision trees in order to make predictions. However, it is worth noting that, in the context of film production, the short time taken to run any of the models is completely inconsequential.

## 6 Conclusion

Based on the results of this research there are a number of findings that can be concluded;

- Basic analysis of a film's script using natural language processing techniques does not appear to produce enough information to make successful predictions about the film's financial performance.
- Use of alternative data about the film, (genre, MPAA rating etc.) appears to produce more accurate predictions, however, these predictions are still quite poor.
- In their current form, neither of these approaches would be able to convince a film producer to employ the use of a predictive model instead of their own expert opinion/ gut feeling.
- The computational time required for the various models is consistently longer for the script data than the movie data by factors ranging from x 1.6 times longer for the naive Bayes model to x 34 times longer for decision tree model.

One possible extension to this work would be to divide each script into acts. As the order of the words in the scripts is not considered by the various prediction methods used in this research, there is no consideration made for the idea that what words go into

making a good first act to a film, may not be the same as the vocabulary used in the second or third act of a screenplay.

One obvious reason why the results of this research were not positive is that movies contain numerous intangible properties that contribute to their success. For example, a film might find itself at the center of some controversy as a result of a dubious connection between one of its leading actors and some unrelated event in his/her personal live. This could affect the financial performance of the film. Another difficulty of trying to predict the success of films based on their script relates to the large variation in how films actually look and feel. As the scripts do not convey which actors will be playing the parts, which composer will score the film or which director will be making the film etc.

The marketing campaign a film receives will also play a major role in how many people go and see the film. An original, initiative marketing campaign could potentially make a dramatic difference to the number of people who go to see a film in the cinema. Two contrasting but effective examples of this are the big budget marketing campaign used by Sony to promote Godzilla, 1998 and the low budget, viral marketing campaign used to promote The Blair Witch Project, 1999 (Dobele et al.; 2005). Whether or not there are any copies of the film leaked onto the Internet could also effect the size of a film's audience. If a film is available to be pirated online, some people will choose this option rather than paying to see the film in the cinema (Danaher and Waldfogel; 2012).

One factor not considered by this research is the film studio responsible for making and releasing the film. Further research could include the studio responsible for producing and distributing the film as a predictor variable in its models. Larger film studios will release their films to wider potential audiences by showing the film at a large number of screens. This will greatly affect the financial performance of the film.

Some advances that could be made to the natural language processing techniques used in this research include the use of n-grams rather than individual words when generating term document matrices. N-grams would allow words to be considered in groups rather than independently. This could potentially result in more insightful data sets being generated from the film scripts.

Another advancement that could be made would be the use of synonym dictionaries, which could be used to reduce the limitations of TF-IDF scores by allowing words with the same meaning to be considered collectively rather than independently.

# References


Appelbaum, E. B. (2004). The consumer price index and inflation - adjust numbers for inflation.
**URL:** *http://www.maa.org/press/periodicals/loci/joma/the-consumer-price-index-and-inflation-adjust-numbers-for-inflation*

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media, *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, Vol. 1, IEEE, pp. 492–499.

*Box office Mojo* (2016).
**URL:** *http://www.boxofficemojo.com/alltime/weekends/month/?mo=08p=.htm*

Breiman, L. (2001). Random forests, *Machine learning* **45**(1): 5–32.

Burgos, M. C., Campanario, M. L., Lara, J. A. and Lizcano, D. (2005). Using decision trees to characterize and predict movie profitability on the us market, *Sahara* **130**: 68–7.

Burr, T. (2013). January is hollywoods very own leper colony.
**URL:** *http://www.nytimes.com/2013/01/20/magazine/how-to-survive-januarys-dearth-of-good-movies.html*

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**: 37–46.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine learning* **20**(3): 273–297.

Danaher, B. and Waldfogel, J. (2012). Reel piracy: The effect of online film piracy on international box office sales, *Available at SSRN 1986299* .

Dobele, A., Toleman, D. and Beverland, M. (2005). Controlled infection! spreading the brand message through viral marketing, *Business Horizons* **48**(2): 143–149.

Elberse, A. and Eliashberg, J. (2003). Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures, *MARKEtiNG SciENcE* **22**(3).

Eliashberg, J., Hui, S. K. and Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts, *Management Science* **53**(6): 881–893.

Eliashberg, J., Hui, S. K. and Zhang, Z. J. (2014). Assessing box office performance using movie scripts: A kernel-based approach, *IEEE Transactions on Knowledge and Data Engineering* **26**(11): 2639–2648.

Ghiassi, M., Lio, D. and Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network, *Expert Systems with Applications* **42**(6): 3176–3193.

Goetzmann, W. N., Ravid, S. A. and Sverdlove, R. (2013). The pricing of soft and hard information: economic lessons from screenplay sales, *Journal of Cultural Economics* **37**(2): 271–307.

Goldman, W. (2012). Adventures in the screen trade, *Grand central publishing* .

Hunter, I., David, S., Smith, S. and Singh, S. (2016). Predicting box office from the screenplay: A text analytical approach, *Journal of Screenwriting* **7**(2): 135–154.

*IMDb* (2016).
**URL:** *http://www.imdb.com/*

Jedidi, K., Krider, R. and Weinberg, C. (1998). Clustering at the movies, *Marketing Letters* **9**(4): 393–405.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features, *European conference on machine learning*, Springer, pp. 137–142.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers, *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., pp. 338–345.

Kim, T., Hong, J. and Kang, P. (2015). Box office forecasting using machine learning algorithms based on sns data, *International Journal of Forecasting* **31**(2): 364–390.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection.

Krauss, J., Nann, S. and Simon, D. (2008). Predicting movie success and academy awards through sentiment and social network analysis.

Landauer, T. K., Foltz, P. W. and Laham, D. (1998). An introduction to latent semantic analysis, *Discourse processes* **25**(2-3): 259–284.

Lantz, B. (2015). *Machine Learning with R*, Packt Publishing Ltd.

Lee, K. J. and Chang, W. (2009). Bayesian belief network for box-office performance: A case study on korean movies, *Expert Systems with Applications* **36**(1): 280–291.

Lee, P. M. (2012). *Bayesian statistics: an introduction*, John Wiley & Sons.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest.

Mestyán, M., Yasseri, T. and Kertész, J. (2013). Early prediction of movie box office success based on wikipedia activity big data, *PloS one* **8**(8): e71226.

MPAA (2015). Theatrical market statistics.
**URL:** *http://www.mpaa.org/wp-content/uploads/2016/04/MPAA-Theatrical-Market-Statistics-2015$_F$inal.pdf*

Mueller, A. (2011). Why movies cost so much to make.
**URL:** *http://www.investopedia.com/financial-edge/0611/why-movies-cost-so-much-to-make.aspx*

Neelamegham, R. and Chintagunta, P. (1999). A bayesian model to forecast new product performance in domestic and international markets, *Marketing Science* **18**(2): 115–136.

Nelmes, J. (2007). Some thoughts on analysing the screenplay, the process of screenplay writing and the balance between craft and creativity, *Journal of Media Practice* **8**(2): 107–113.

Paice, C. D. (1994). An evaluation method for stemming algorithms, *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Springer-Verlag New York, Inc., pp. 42–50.

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries, *Proceedings of the first instructional conference on machine learning*.

Rish, I. (2001). An empirical study of the naive bayes classifier, *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3, IBM New York, pp. 41–46.

*Screenplays For You* (2016).
**URL:** *https://sfy.ru/*

Sharda, R. and Delen, D. (2006). Predicting box-office success of motion pictures with neural networks, *Expert Systems with Applications* **30**(2): 243–254.

Simonoff, J. S. and Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers, *Chance* **13**(3): 15–24.

Statista (2016a). Global filmed entertainment revenue 2015-2020.
**URL:** *https://www.statista.com/statistics/259985/global-filmed-entertainment-revenue/*

Statista (2016b). Statistics and facts about the film industry.
**URL:** *https://www.statista.com/statistics/259985/global-filmed-entertainment-revenue/*

*The Daily Script* (2016).
**URL:** *http://www.dailyscript.com/*

*The Internet Movie Script Database* (2016).
**URL:** *http://www.imsdb.com/*

*The Numbers* (2016).
**URL:** *http://www.the-numbers.com/*

Vogel, H. (2014). Entertainment industry economics:a guide for financial analysis., *Cambridge University* .

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Zhang, W. and Skiena, S. (2009). Improving movie gross prediction through news analysis, *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, IEEE Computer Society, pp. 301–304.

Zufryden, F. (2000). New film website promotion and box office performance, *Journal of Advertising Research* **40**(1-2): 55–64.

# A

# Tables of Results

Table 3: 922 movie data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.62 | 0.49 | 0.39 | 0.43 |
| Random Forest | 0.62 | 0.51 | 0.43 | 0.47 |
| Naive Bayes | 0.63 | 0.55 | 0.12 | 0.20 |
| SVM | 0.62 | 0.51 | 0.44 | 0.47 |

Table 4: 922 script data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.52 | 0.37 | 0.39 | 0.38 |
| Random Forest | 0.62 | 0.35 | 0.03 | 0.04 |
| Naive Bayes | 0.62 | 0.39 | 0.03 | 0.06 |
| SVM | 0.53 | 0.38 | 0.40 | 0.38 |

Table 5: 922 LSA data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.54 | 0.40 | 0.46 | 0.42 |
| Random Forest | 0.58 | 0.40 | 0.31 | 0.33 |
| SVM | 0.47 | 0.34 | 0.59 | 0.42 |

Table 6: 922 movie data, no release month

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.62 | 0.51 | 0.30 | 0.37 |
| Random Forest | 0.63 | 0.53 | 0.34 | 0.42 |
| Naive Bayes | 0.63 | 0.10 | 0.00 | 0.01 |
| SVM | 0.52 | 0.39 | 0.50 | 0.44 |

Table 7: 741 script data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.59 | 0.31 | 0.31 | 0.31 |
| Random Forest | 0.70 | 0.00 | 0.00 | 0.01 |
| Naive Bayes | 0.69 | 0.00 | 0.00 | 0.00 |
| SVM | 0.62 | 0.23 | 0.14 | 0.17 |

Table 8: 741 movie data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.67 | 0.42 | 0.30 | 0.35 |
| Random Forest | 0.66 | 0.40 | 0.30 | 0.36 |
| Naive Bayes | 0.70 | 0.00 | 0.00 | 0.00 |
| SVM | 0.66 | 0.41 | 0.27 | 0.32 |

Table 9: 741 LSA data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.58 | 0.33 | 0.39 | 0.34 |
| Random Forest | 0.63 | 0.27 | 0.16 | 0.12 |
| SVM | 0.50 | 0.29 | 0.49 | 0.35 |

Table 10: 741 movie data, no release month

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.70 | 0.49 | 0.11 | 0.18 |
| Random Forest | 0.69 | 0.43 | 0.13 | 0.21 |
| Naive Bayes | 0.70 | 0.00 | 0.00 | 0.00 |
| SVM | 0.53 | 0.29 | 0.41 | 0.33 |

Table 11: 663 script data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.59 | 0.69 | 0.67 | 0.68 |
| Random Forest | 0.65 | 0.65 | 0.98 | 0.78 |
| Naive Bayes | 0.64 | 0.65 | 0.97 | 0.77 |
| SVM | 0.59 | 0.65 | 0.78 | 0.71 |

Table 12: 663 movie data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.56 | 0.68 | 0.63 | 0.65 |
| Random Forest | 0.59 | 0.67 | 0.72 | 0.70 |
| Naive Bayes | 0.65 | 0.65 | 1.00 | 0.79 |
| SVM | 0.57 | 0.66 | 0.72 | 0.69 |

Table 13: 663 LSA data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.56 | 0.70 | 0.60 | 0.62 |
| Random Forest | 0.60 | 0.69 | 0.75 | 0.68 |
| SVM | 0.45 | 0.67 | 0.30 | 0.38 |

Table 14: 663 movie data, no release date

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.60 | 0.67 | 0.77 | 0.71 |
| Random Forest | 0.61 | 0.66 | 0.81 | 0.72 |
| Naive Bayes | 0.65 | 0.65 | 1.00 | 0.79 |
| SVM | 0.49 | 0.63 | 0.52 | 0.56 |

Table 15: Comedy script data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.55 | 0.51 | 0.48 | 0.49 |
| Random Forest | 0.55 | 0.57 | 0.35 | 0.44 |
| Naive Bayes | 0.56 | 0.54 | 0.33 | 0.41 |
| SVM | 0.48 | 0.44 | 0.48 | 0.44 |

Table 16: Comedy movie data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.63 | 0.63 | 0.49 | 0.53 |
| Random Forest | 0.60 | 0.57 | 0.55 | 0.55 |
| Naive Bayes | 0.53 | 0.49 | 0.22 | 0.30 |
| SVM | 0.63 | 0.61 | 0.55 | 0.57 |

Table 17: Drama script data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.52 | 0.41 | 0.42 | 0.41 |
| Random Forest | 0.55 | 0.48 | 0.17 | 0.16 |
| Naive Bayes | 0.60 | 0.55 | 0.18 | 0.25 |
| SVM | 0.55 | 0.44 | 0.36 | 0.39 |

Table 18: Drama movie data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.62 | 0.55 | 0.47 | 0.50 |
| Random Forest | 0.62 | 0.55 | 0.52 | 0.52 |
| Naive Bayes | 0.61 | 0.53 | 0.11 | 0.17 |
| SVM | 0.63 | 0.56 | 0.51 | 0.53 |

Table 19: Action script data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.60 | 0.29 | 0.39 | 0.33 |
| Random Forest | 0.73 | 0.00 | 0.00 | 0.00 |
| Naive Bayes | 0.73 | 0.00 | 0.00 | 0.00 |

Table 20: Action movie data

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 0.71 | 0.34 | 0.23 | 0.25 |
| Random Forest | 0.73 | 0.34 | 0.28 | 0.29 |
| Naive Bayes | 0.75 | 0.00 | 0.00 | 0.00 |

Table 21: 922, 3 outcomes

| Model | Accuracy | Cohen's Kappa |
|---|---|---|
| Decision Tree | 0.49 | 0.25 |
| Random Forest | 0.54 | 0.26 |
| Naive Bayes | 0.52 | 0.31 |

Table 22: 922, 3 outcomes, no release month

| Model | Accuracy | Cohen's Kappa |
|---|---|---|
| Decision Tree | 0.51 | 0.29 |
| Random Forest | 0.51 | 0.23 |
| Naive Bayes | 0.54 | 0.26 |

Table 23: 922, 5 outcomes

| Model | Accuracy | Cohen's Kappa |
|---|---|---|
| Decision Tree | 0.40 | 0.21 |
| Random Forest | 0.42 | 0.23 |
| Naive Bayes | 0.49 | 0.25 |

Table 24: 922, 5 outcomes, no release month

| Model | Accuracy | Cohen's Kappa |
|---|---|---|
| Decision Tree | 0.40 | 0.21 |
| Random Forest | 0.42 | 0.22 |
| Naive Bayes | 0.49 | 0.25 |

Table 25: Computational time in seconds, 922 films

| Model | Script data | Movie data |
|---|---|---|
| Decision Tree | 47.83 | 1.40 |
| Random Forest | 165.79 | 22.77 |
| Naive Bayes | 2.51 | 1.54 |
| SVM | 3.04 | 33.13 |