INFLUENCE OF SOCIAL MEDIA ON PERFORMANCE OF MOVIES

A Thesis presented to the Faculty of Graduate School
University of Missouri

In Partial Fulfillment of the Requirements for the degree
Master of Science

by
FNU SHRUTI

Dr. Wenjun Zeng, Thesis Advisor
MAY, 2014

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled

INFLUENCE OF SOCIAL MEDIA ON PERFORMANCE OF MOVIES

Presented by FNU SHRUTI

A candidate for the degree of Master of Science

And hereby certify that in their opinion it is worthy of acceptance.

Dr. Wenjun Zeng

Dr. Jianlin Cheng

Dr. Tony X Han

ACKNOWLEDGEMENT

I would like to thank my thesis advisor, Prof. Wenjun Zeng, for his invaluable advices, support and help during my study in University of Missouri. Without his constant encouragement, motivation and inspiration, this thesis would not be possible. I would also like to thank my other committee members, Dr. Jianlin Cheng and Dr. Tony Han, for their valuable comments and suggestions for this work.

I would like to extend my sincere gratitude to Dr. Suman Deb Roy for his immense support, direction and guidance towards the process of completion of this work. I would also like to thank my colleagues in the mobile networking and multimedia communications lab, for their friendship and help.

Finally I express my deepest gratitude goes to my family, who has provided me with unconditional love and support throughout the period of my studies at Mizzou. I would like to thank my parents for raising me to who I am. Their advice has always been the best guidance in my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The Motion Picture Industry has often been referred to as “Land of Wild guesses and hunches”. As films attempt to reciprocate to the advancing expectations of people, it has led to exponential rise in the risk associated with making any film. We thus feel that it is important to analyze and determine the factors which could affect the financial performance of movie. The performance of movies in terms of revenue depends on many factors such as its production studio, genre, script quality, pre-release promotion etc, - all of which are traditionally used to estimate its potential success at the box office. Recently however, the “Wisdom of Crowd” and social media have been acknowledged as a strong signal in understanding consumer behavior towards media. In this thesis, we capture socially generated meta-data mined from the social media and multimedia sites such as number of likes on Facebook page for the movie, follower count of actors on Twitter and number of likes on trailer of the movie on YouTube and study their influence on box-office performance and profitability of movies. We try to study the influence of social media signals we collect in classifying movie profitability. We identify the performance of movies by classifying them into 3 profitability classes. We test well-known machine learning algorithm for this purpose and compare the results to find which algorithm is the most suitable for our data and the problem scenario. We also do another comparison study to test which social media signal is the strongest predictor of movie profitability. Our result shows that various social media signals have varying yet significant impact in predicting the performance of movies. Our research also reveals that popularity of actor depicted through follower count on Twitter is most relevant to the success of movie at theaters, and Facebook ‘like’ signal has noise which impedes its analytical credibility.

CHAPTER 1

INTRODUCTION

Films have always been the most cherished source of entertainment through all times. Film reciprocates the advancing expectations of the public which has led to exponential rise in the risk associated with making any film. There is inherent unpredictability in terms of estimating the revenue that can be made out of it. Hence, the challenge the motion picture industry faces is to be able to estimate and predict the revenue that will be generated by a given movie owing to the huge investment it involves. The intensity of problem increases when we realize there are multitude factors which impact the revenue of the movie.

Some factors are traditionally related to the movie such as MPAA rating, budget, opening theaters while other are socially generated promotions through signals such as number of likes on Facebook page for the movie, follower count of actors on Twitter and number of likes on trailer of the movie on YouTube. Such a scenario gives us an opportunity to explore socially generated meta-data extracted from social media and multimedia sites and assess their impact on the performance of movies.

Previous research has tried to solve this problem; however it does sub-optimally by choosing traditional factors such as advertisement budget, star cast, production house, movie script, sequel, MPAA rating, number of opening theaters etc. Yet, there is a demand for more accurate classification model.

The tales of two movies, Jack Reacher and Rock of Ages, released in the year 2012 were our motivation towards delving more into this field of research and unraveling the structures which have been ignored in the pasts. Table 1 lists the traditional factors like

budget, MPAA rating, number of theaters, genre etc. We see that there is no significant difference in the values of these features which could lead to noteworthy difference of $175 million in the gross values. This scenario indicates that socially generated data could add more information about a movie's potential at the box-office beyond traditional attributes and so would be interesting for exploration in the analysis of the factors which influence a movie's performance at box-office.

TABLE 1- Traditional factors of Rock of Ages and Jack Reacher

TRADITIONAL VARIABLES	**ROCK OF AGES**	**JACK REACHER**
Distributor	Warner Bros.	Paramount
Release	June 15 2012	Dec 21 2012
Run Time	2 hrs 3 min	2 hrs 10 mins
MPAA Rating	PG-13	PG-13
Genre	Musical	Crime Drama
# of Theaters	3470	3352
Opening Weekend	$14 million	$15 million
Worldwide Gross	***$59 million***	***$216 million***

Over the past few years, Social Media is being enthusiastically used in exploring this domain of research further. There is a lot of socially generated meta-data generated owing to the huge amount of movie promotions being done on social media. Such meta-data reflect users' reaction towards the movie and could be used to capture their approval instances. Thus, social media seems to offer an innovative paradigm to address such problems. Consequently, traditional methods have definitely come under scrutiny. A number of interesting lines of recent work [4][5][6][7] have pursued the pre-release financial movie success problem using a variety of social media data such as Google and YouTube trailer search volume, critic reviews, blogs, tweets, Wikipedia activity level. Moreover, social media signals have been used in related researches such as in flu predictions, election predictions, music recommendation system etc, which gives some references for leveraging social media to do analysis in our work.

Despite these developments, there is still room to add value which could aid the accuracy of the classification and prediction problems. Let us spend some time in understanding the major constraints with the ongoing research scenario.

First, our generic understanding suggests before tackling a prediction problem, it is important to address the problem of relevance between the source and target variables and perform successful classification. So, in our study, we try to identify the social signal which could be the most relevant to performance of movie with an understanding that upon availability of the social signal on a timeline, it could act as a reliable source of data with enhanced predictive capacity.

We analyzed a few works [8][4] which have used buzz on Twitter in form of tweets and Wikipedia activities to develop predictive models for the box-office revenue of the movie. Data was in the form of volume of tweets about the movie or the number and frequency of edits on Wikipedia page for the movie. Deploying these variables, predictive models were constructed. It is evidently seen that the performance of the models significantly degraded for movies which fell in low and medium popularity range. We attribute such degradation in the performance of the model to the choice of social media signal made. It is understood that less popular movies do not get much audience attention and so socially generated data pertaining to those may not be in significant volume suitable for study. Still, we strongly believe that there are some social signals which could be truly indicative of the user's attention and might be helpful in generating non-skewed classification/prediction model, may be, due to the higher number of active users. In our study, we chose to explore meta-data generated on the most popular social media venues and capture signals such that they could be easily extrapolated as general opinion owing to the huge number of its daily regular users.

Upon investigation, we observe that current research suffers from the following limitations:

 Social media signals such as tweet, blogs and reviews lack worldwide applicability due to language difference.

 Publicly available meta-data on social multimedia sites like YouTube still remains unexplored in addressing the problem.

 Star popularity in influencing the performance of movies still stands in debate.

 A comparative model to compare the relevance of socially generated meta-data from media and multimedia sites remains absent.

We formulate two concrete tasks that address the realms of the issues discussed above.

 Determine *whether social media signals improve classification of movies by profitability*.

 Do a *comparative analysis to test the amount of relevance that different social media signals have on the performance of movies*.

Considering the constraint with availability of public data on social media sites, we perform classification and consider it as an initial stage of a prediction problem. Our research uses purely statistical attributes and so is independent of any language based analysis like sentiment analysis. Moreover, our classification is performed on 532 movies which surpass other relevant work in terms of the size of data under investigation.

The main contributions of our work are as follows:

 We show that social media signals improve the classification of movies when used in combination with some traditional attributes. Traditional variables are less dominant in driving the box-office performance of movies.

 We discover that popularity of movie casts (actors) depicted through Twitter data is the strongest social media signal to box office performance of movies.

 Facebook ‘like’ signal is noisy and do not classify movies accurately.

 The Support Vector Machine (SVM) algorithm gives us the best classification accuracy, in comparison to Boosted trees and Adaptive Bagged trees algorithm, in terms of F-Score for different profitability classes and the average value.

In Chapter 2 of this thesis, we describe related work in this domain of research where we discuss related work using traditional features used to study this problem followed by studies which used social media signal for the same. We shall also be describing the role that social networking sites play in influencing the revenue a movie makes. In Chapter 3, we describe in details the dataset we have, the sampling and the reasons which influenced the selection of the social signals we use. Following this, in Chapter 4, we discuss about experiments, results and our analysis when we try to determine the importance of social media in influencing the box-office revenue of movies and the algorithm which suits our dataset and the problem scenario. In Chapter 5, we again discuss the experiment, result and analysis when we try to determine the most relevant social signal in influencing the revenue of movies. In Chapter 6, we conclude with the conclusions we made from our study and the future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

In this chapter, we focus on the technical background required prior to diving deeper into this thesis. We shall also mention key state-of-the-art research results so that it is easy to distinguish our contributions from existing previous work. Since we deal with different kinds of data, we shall focus on the different sources of data extraction and their contribution in understanding this domain of research. We shall then focus on the technologies that have been and are being used to discover pattern and gain knowledge from the data. These mainly employ data mining and machine learning techniques.

2.1 Learning from Mainstream Media:

By 1980s and 1990s, several researchers had begun to work proactively in the field of analyzing movies performance and predicting its success in terms of the revenue it generates Their efforts have helped establish and define traditional variables used in this field of study. Many interesting lines of works have pursued their research using genre, scripts, MPAA ratings, season of release, runtime, directors, studios, budget, awards and nominations and many other as input variables to predict opening weekend revenue or classify movies based on their box office performance.

Under the umbrella of traditional factors, we see that previous research has laid focus on the following features/factors in developing predictive/classification models to analyze the financial success of movies. Some of the studies we discuss below also depict efforts made by the early researchers to determine some of the most profitable venues for advertising and pre/post release promotion.

- **Production Banner, Budget and Team**

We see that early research work [2], one of the earliest works in this domain, hypothesizes that greater production budget value being built into the film turns to lead towards greater popularity and quality of the finished work. Such popularity generally leads to larger revenue made by the movie at the theaters. Thus, the author believes production budget to be a strong determinant towards film's quality and gross. A team evaluation approach proposed by Shugan in [16] says that past performance of the production team could be one of the strong predictors towards estimating the revenue the most could make on release.

- **MPAA Rating**

The **Motion Picture Association of America's (MPAA) film-rating system** is used in the United States (US) and its territories to rate a film's suitability for certain audiences. We see that early works have paid significant attention to the rating of the movie and have investigated its predictive capacity largely. Though Litman in his earliest work [2] emphasizes on PG being the most desirable rating as it could reach maximum audience, we see that with coming years, there were varying conclusions made by other researchers. In 2004, Leenders and Elaishberg [17] found that PG-13 has gradually become a more common rating and that parental guidance for a particular movie changes across countries and so might not be very appropriate to be considered globally.

- **Time of Release**

We also get to see that holiday seasons are cited to be the best times to release a film as well a time for heavy competition. We see in a work by Krider and

Weinberg [18] that movies try to avoid release in the season for movies which are intended for similar audience. Chisholm in [19] models the competition between movies' release timing as a war of attrition. She suggests studios play a complicated game while choosing time of release. Sochay [20] also finds out that during festive or holiday seasons, bigger box-office hits might have ripple effect on other movies released around the same time.

- **Awards and Nominations**

  Litman [2] argues that critical acclaim lends momentum to a film's theatrical success whereas unfavorable critical press can have a decelerating impact on a film's theatrical success. Litman [2] saw the awards as a way to accelerate box-office gross for a short time period. Later we see that other researchers have carefully examined the potential of different awards towards influencing movie gross and found out that not all movies bear same effect. In particular, Smith and Smith [21] examine specific awards that are the best indicators of a film's revenue. Award nominations are released post-release; therefore, to approximate pre-release impact, we could take a count of the number of awards the cast and crew of the film have been nominated for and/or won in past films.

- **Number of theaters upon release**

  A common conjecture by researchers is that the more the number of theaters the movie releases into, the more revenue it generates. There have been a few studies which conclude that the longer the movies are on screen in theaters, more is the gross they make.

- **Hollywood Stock Exchange**

Another interesting new method involves the use of stock market simulations. Some marketing researchers have shown that such 'predictive' markets can generate, at an early stage, valuable insights into the likely success of motion pictures ([22][23][10]). Spann and Skiera in [10] show that data obtained using HSX (Hollywood Stock Exchange), when incorporated into a conventional regression model, leads to a significant improvement in opening weekend forecasts. One possible reason for why virtual stock markets are helpful in assessing demand stems from a key observation about movie consumption – moviegoers appear heavily influenced by others' opinions and choices. 'Others' could refer to friends and acquaintances, critics and other opinion leaders, as well as the market as a whole.

- **Sequels**

  There has always been a common conjecture around the studios that sequel of a hit movie would make more profit. Interestingly, we see that some researchers deny the conjecture. Marc Schmuger, Vice Chairman at *Universal Studios*, commented in this regard: "It's a complex equation that figures in determining whether the sequel is capable of capturing the same level of excitement as the original" (Variety 2003d). Interesting in this regard, Sood and Dreze in [24] , who consider movie sequels as brand extensions and focus on the role that their titles play, find that a sequel with a numbered title (e.g., *Daredevil 2*) may have a less favorable evaluations than a sequel with a more descriptive title (e.g., *Daredevil: Taking it to the Street*).

- **Star Power**

Several researchers have studied the effect of star power. Most studies consider star power as one of the covariates in a regression model with box office performance as the dependent variable [2][25][20][26][27]. Focusing solely on the role of stars, Albert [28] empirically shows that stars serve as the most consistent 'markers' for successful films which, he argues, explains their power in Hollywood. However, also using a probability modeling technique, De Vany & Walls in [29] conclude that audiences make movies hits, and "no amount of 'star power' or marketing can alter that". In another study on the role of stars, Ravid [3] finds no correlation between star participation and film revenues or profitability, which is consistent with the view that stars capture their 'economic rent'. Overall, existing evidence on the extent to which stars drive box office performance is mixed, and more research is needed to resolve this debate. With a unique method to measure star power, Brewer et al. in [30] used the Harris Poll to measure the top ten movie stars, and then combined this with the People's Choice Awards for Best Motion Picture Actor/Actress, Male/Female TV Performer and another poll used to measure star popularity. The stars that were present on all three polls for the years studied were then included in the final list of sixty-six.

- **Genre**

In an attempt to understand this problem from a different angle, some researchers have also used the concept of genomes: semantic meta-data about the movie which could range from fine-grained semantics such as mood, plot, audience type, praise, style and whether it is based on a book or not to more traditional classes such as, genre, musical score, flags of violent content, Oscar-winners etc. These set of

semantic features for a movie is called its genome [5]. Alternately, each semantic feature (e.g., mood) represents a gene. Their study identified 4 communities of genes which have positive impact on the revenue of movie and five communities of genes which have negative impact with an accuracy of 71% in predicting high profitability movies 0.

2.2 Learning from Social Media

Advancement in information technology, reduction in storage cost, and development in the field of machine learning has made user generated data a much talked about variable in present times. In this age of huge prevalence of individuals connected through social media and “word of mouth” been recognized lot of times in varied field of research, traditional factors have recently come under scrutiny in this domain of study. Social Media gives ordinary people the power to be content creators and information disseminators. This information is embedded in multimedia shared across social networks, containing valuable indications about various facets of human life - what captures our attention, our sharing biases and digital traces we abdicate.

Figure 1- Depicting various forms of user generated content on different social media platforms

Social media has become a disruptive platform for addressing many multimedia problems elegantly [31]. It has penetrated every realm of business and academia (marketing, advertising, journalism, broadcast, stock markets etc.) and its existence is ubiquitous. Moreover, remarkable insights can be extracted from social media. For example, real-time social data is being utilized in a number of scenarios - from visualizing political activity and flu outbreaks [35][32], forecast and prediction to sentiment detection [34] and emergency advisory systems [33].

Social media has also largely affected existing models of communication and information retrieval. Social networking use is steadily increasing among the key demographic of teen and young adult moviegoers. The continuing popularity of social media has opened direct channels to potential customers that were not previously available to movie marketers. Sharing sites such as YouTube, Facebook, and Twitter have increased the spread of information to lightning speeds.

Promotion on Social Media

According to Variety, the box office is pacing higher for Summer 2013 than the all time highest from back in 2011. But movie marketers face more challenges than ever before. Consumers are adopting new social technologies with increasing speed. IAB (Interactive Advertising Bureau) reports that 42% of US smart-phone users under the age of 35 now check their favorite social media apps before deciding on which movie to see. The marketplace is changing, and while most coverage has explored the best overall digital movie programs, the focus of the following observations relate to new trends in film marketing.

FACEBOOK: This social media site reaches 142M unique users a month. It is the #1 social website reaching 78% of internet visitors. It is also the anchor of social media strategies for most films. Movie marketing teams launch their Facebook pages sometimes as early as the day they announce a film, using them to promote other programs in different social channels such as live chats on Twitter, channels on Spotify, or trailers on YouTube as seen in Figure 2. The key is bringing all this back to a social channel with the broadest reach to get these new movies out to the masses.

The newest trend with Facebook movie promotions is purchasing targeted ads on mobile. Over 70% of demographic having age between 18 to 34 access Facebook via mobile. As moviegoers spend more time on social via their mobile devices, and as most movie viewing decisions are made by checking movie ads and updates, this is the best way to capture these consumers in market.

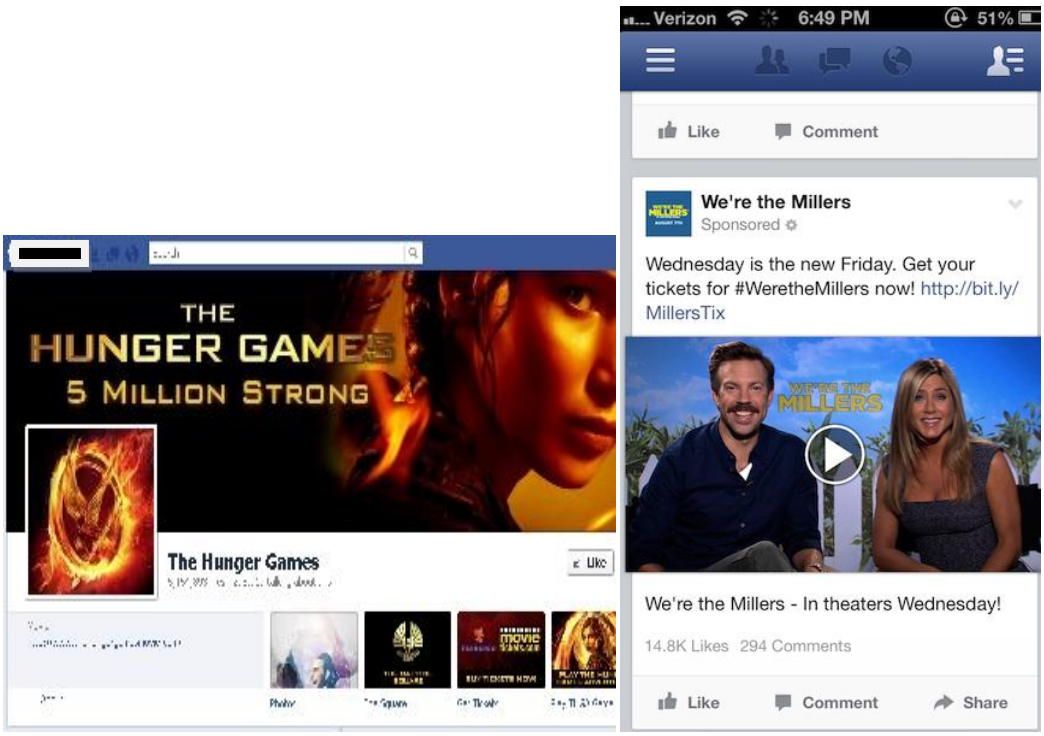

Figure 2- Promotion on Facebook through Facebook page and advertisements (sponsored/purchased) for the movie 'The Hunger Game' and 'We're the Millers'

TWITTER: It is the #2 social network/micro blogging platform reaching 28M unique users a month. Television rules the Twitter-sphere, and while many movies maintain a presence on Twitter, very few film pages reach over 1M followers. Entertainment marketers leverage in-tweet media to enable mass distribution of trailers. Brands have also started to include hashtags in their TV advertisement and other marketing materials to encourage conversations. Twitter has become one of the most accessible platforms for actors seeking popularity. It is an emerging trend to measure star popularity with the number of followers they have on Twitter. Every year most prominent magazines and websites release the most popular actor list using the number of followers they have on Twitter. Figure 3 below shows a list of the most popular people in the world using Twitter followers as the measure. We see except for Barack Obama, 3 out of 4 world's most

popular entities with maximum followers are actors and are related to the entertainment industry.

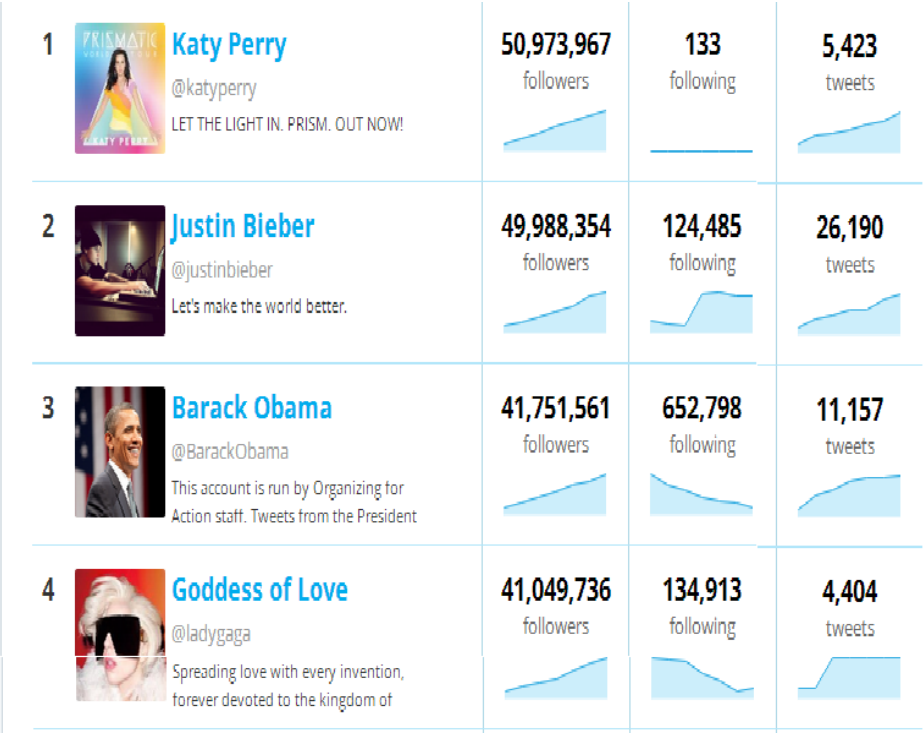

Figure 3- Depicting top 4 popular people in the world in 2013 by their follower count on Twitter

Twitter provides people a way to reach out to their favorite actor and keep updates about the latest happenings of their life, thus, stay connected. Twitter also gives the stars an easy way to reach out to their fans and keep giving their fans updates about their latest projects. This may act as a means to keep the interest about their upcoming films alive. Past researchers have used various measures to tap buzz around actors/movies. Traditionally, researchers have used Harris Polls and magazine Polls to capture the stardom of actors. Also, some studies have used star ratings through IMDB and Rotten Tomatoes. Recently, social media has started to be used to measure actor and movie buzz. A lot of studies have used buzz about a movie on Twitter and predict the performance of movies. With some

considerable successes that early studies got using the above-mentioned metrics, we thought it would be interesting to explore more about signals which could reflect popularity of actors. After careful analysis of the different venues which could act as a reliable source for measuring an actor's popularity, we chose to tap the *follower count for actors on Twitter* and compared it with other social signals. We chose Twitter over other domains as we know that Twitter is the most popular platform used by famous personalities and a large number of people in the world.

YOUTUBE: Whether it's general curiosity or full engagement with a film, moviegoers are constantly searching for information. Online engagement through search allows for the ability to interact with moviegoers in real time, giving them the chance to ask questions and receive immediate feedback. Trailers are one of the most influential sources throughout the decision process to see a movie. In fact, it was found that trailers are the most searched for category of information upon discovery of a new film [5]. Trailer searches, whether on Google or YouTube, signify strong intent -- searchers are actively seeking a sample of the film. Thus, trailer-related search query volume holds strong predictive power. In a recent survey, it was found that most moviegoers learn about a film four weeks in advance. Similar to trailer-related Google searches, title-related searches on YouTube have the highest predictive power four weeks from release date (R^2 (Coefficient of determination) = 55%) -- even stronger than the predictive power of release week searches [5]. With YouTube having launched Trailers, an exclusive section for showcasing the Hollywood movie trailers in HD quality, trailer related data has become an interesting point to study as shown in Figure 4. Even though there are many blog and sites that offers trailers, YouTube Trailers organizes the videos in a neat

way. We can browse through the categories like most popular, latest, opening soon, and those currently in theaters.

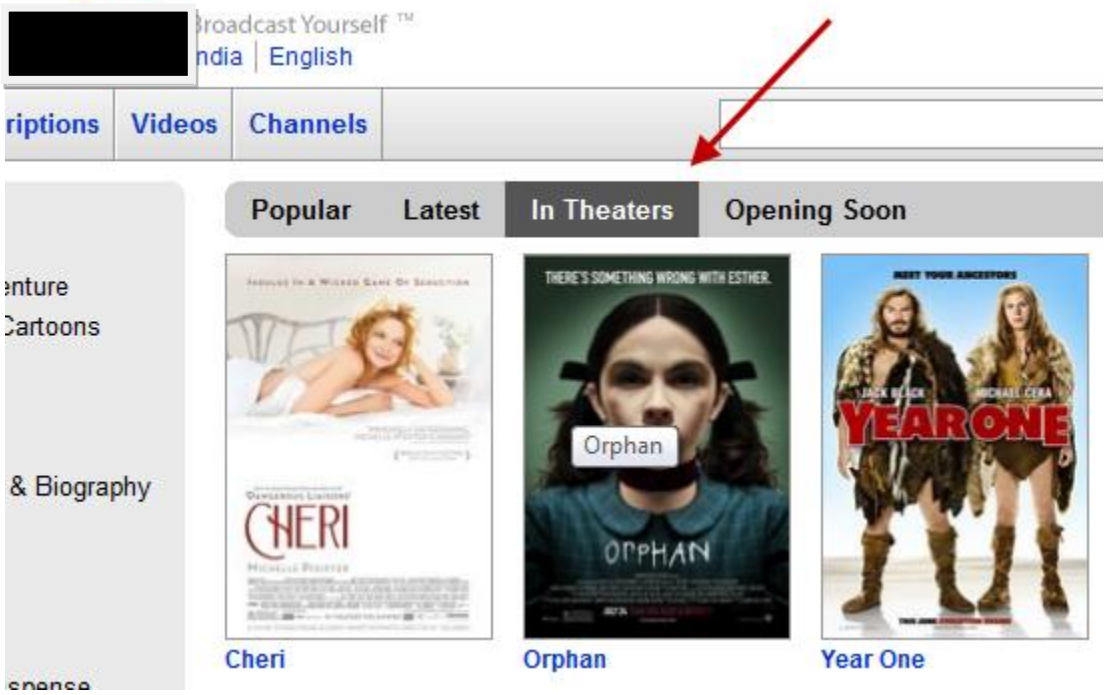

Figure 4- Dedicated channels on Youtube promoting movies

2.3 **Related Works**

Many previous studies have shaped the goal of our work. Ishii et al. developed a mathematical framework for the spread of popularity of the movie in society [36]. Their model considers the activity level of the bloggers estimated through number of weblog posts on particular movies in the Japanese Blogosphere as a representative parameter for social popularity. Similarly, other researchers have developed models taking the activity level of editors on Wikipedia as a popularity parameter [4]. On a criticizing note, Mashine and Glances in their work analyzed the sentiment of Weblog stories on movies, and pointed out that the correlation between pre-release sentiment and sales is not at an adequate level to build up a predictive model [37]. Going with the trend, many other

studies have tried to capture popularity variable through movie and star buzz through tweets or pre release hype created on the Internet through advertising and marketing. On one hand, a few studies [8] advocate that social media buzz have positive impact on box office returns, some others [7] advocate that buzz in the form of tweets do not necessarily relate to box office revenue. Star Power has been one amongst the features on which recent researches have focused much on. There is a wide array of research through years [9][20][3] which have been debating on star power. In light of the above work, we try to investigate star popularity/power factor as well through follower count of the actors of movie and compare it with other social media variables which depict popularity of the movie as a whole. Hence, in our study, we try to find the most relevant social multimedia signal which could influence the box office returns of movies.

CHAPTER 3

DATA DESCRIPTION, SELECTION AND SAMPLING

3.1 Data Description

Dataset with traditional attributes: In order to apply our experimentation to real world data, we used the dataset that was released by MPAA after the Oscar Academy Award 2011. We took reference from previous studies in order to eliminate statistically insignificant features. Previous researches have pointed out that variables such as genre, runtime, MPAA ratings, Studio and Oscar nomination and awards do not correspond significantly to box office performance of movies [11]. Hence, after careful examination, we eliminated some of those variables. Thus, we used the following combination of traditional attributes for 532 movies as shown with an example in Table below.

Social Multimedia Signals augmented into the dataset: With an intention to discover the social media signals that potentially possess stronger correlation with the profitability of a film, we identify signals which reflect audience approval from different social media domains. Note that the type of data in each domain may be different, e.g., Twitter is a social stream whereas YouTube is a social video publishing website. Most of social media buzz around a movie that have been captured for this study are before its release or during the first 1-2 weeks.

. We chose the following social media signals for our research.

- *Facebook like count on movie pages*, motivated by a recent study which illustrates that Facebook is probably the most disruptive context in which we can see the

pages, statuses etc powered by the Like button. Theories of social influence suggest that the like count might influence people's decision making, in our case the decision to buy a ticket to watch movie in theater [9].

- *Follower count of the actors in Twitter*, motivated by many recent studies which advocate that popularity of individuals like directors/actors of the movie have direct relationship with box office performance of movies. Buzz about stars has been used to study such problem scenario by some researchers. In [9], the author uses star buzz using the STARMeter tool of IMDB to predict movie performance with 3SLS (3 stage least square) non-linear regression and suggests that buzz around movie stars may impact its performance. This gives us motivation towards exploring more about signals which could reflect popularity of actors. After careful analysis of the different venues which could act as a reliable source for measuring an actor's popularity, we chose to tap the *follower count for actors on Twitter* and compared it with other social signals. We chose Twitter over other domains as we know that Twitter is the most popular platform used by famous personalities and a large number of people in the world. Twitter offers an ideal venue for users to follow their favorite actors and can be tapped to assess the popularity of the actors through follower counts [11]

- *YouTube like count on the official trailers of the movies*, motivated by a recent study by Google Team suggesting Trailers remain one of the most influential sources throughout the decision process to see a movie and trailer-related search

query volume holds strong predictive power [5]. We use likes to capture approval instances.

Though the data got from these sites can be criticized for their "representativeness" being unclear and noisy, but these sites are vast storehouses of information pertaining to individuals connected through social media like Facebook, YouTube, Twitter etc. Though follower count of an individual's Twitter account is being criticized for its unaccountability towards assessing his popularity due to attacks by bots or fake accounts, we make an attempt to fairly avoid the accounts that are following many other accounts, but followed by only a few of them and taking only those accounts which have been verified by Twitter with a 'checkmark' as suggested in [15].

Table 2- the set of traditional attributes along with social media signals for the movie Battle: Los Angeles

Name	Battle: Los Angeles
Rotten Tomatoes score(critic)	35
Audience Score	50
Budget($M)	70
Worldwide Gross($M)	211.82
Box Office Average of Opening Weekend($M)	10.4
Number of Theaters in Opening Weekend	3417
Facebook Likes	868507
Twitter Followers	551.5
Youtube Likes	3700

3.2 **Data Sampling and Selection**

The major reason for selecting the specific social signals this thesis discusses lies in their applicability in the problem domain. Researchers have not studied these signals for the

movie popularity prediction problem before. Also, selection of the above meta-data has partly been supported by the APIs data features/call these products offer. Meta-data extracted from these sites gives us responses from a large number of users which can be extrapolated as general opinion owing to the fact that platforms like Facebook and Twitter claim to have the highest number of active users across the globe while YouTube is one of the most popular social multimedia sites.

We used the following APIs to retrieve the values of the signals for the movies. We parsed the output in JSON format. All the APIs have been implemented in java.

We used Facebook "Graph API" with '/page' reference to retrieve the number of likes on the movie page on Facebook.

We used YouTube's "Data API v3.0" to retrieve the number of likes on the official trailer video of the movies.

We used Twitter's "Lookup API v1.0" to retrieve the follower count on verified user accounts for actors.

It is important to tap the information available here in order to gain insight of the popularity of movies amongst audience and customers, to comprehend the feelings of moviegoers and predict the number of ticket sales.

Also, in order to give a better insight of the dataset, Figure 5 below shows the class distribution for the dataset.

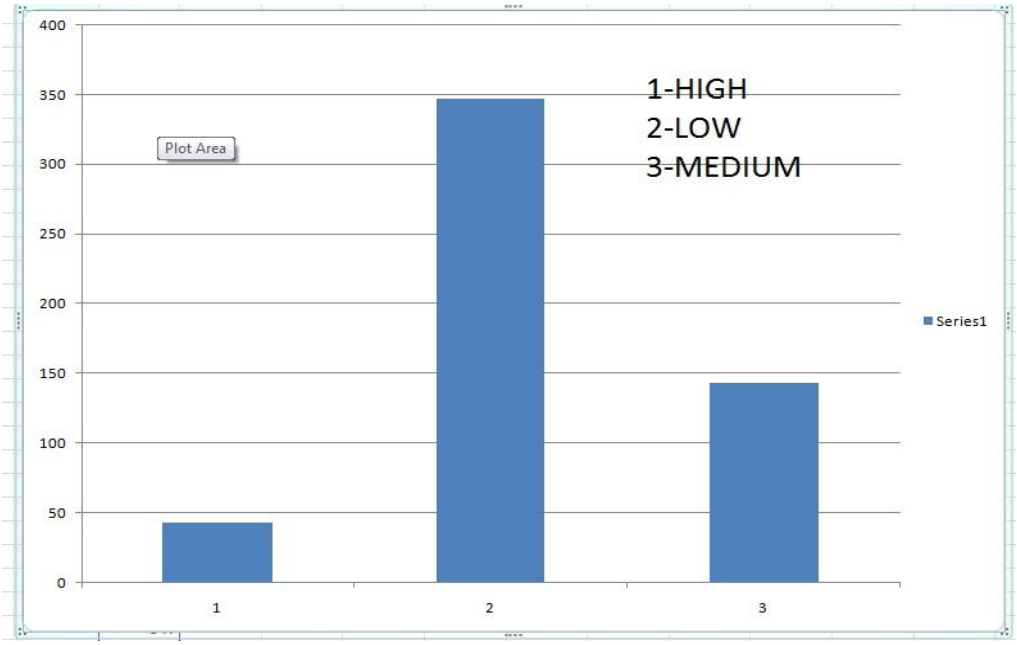

Figure 5- Class distribution of the sample collected

CHAPTER 4

DETERMINE POTENTIAL SUCCESS OF MOVIE WITH SOCIAL MEDIA SIGNALS

4.1 Introduction

With such high influence of social media on the daily life of people, we think it would be interesting to study user response towards a media (here, movies) by capturing their sentiments expressed on social media platforms like Facebook, Twitter and YouTube. With the advent of social media and it having become such an integral part of human life, peer influence in the process of decision making is also a commonly observed behavior. In this study, we try to tap approval instances of users for particular movies from very popular social media platforms mentioned above.

We, thus, try to study the influence of social media signals in classifying movie profitability. We identify the performance of movies by classifying them into 3 profitability classes: *low*, *medium* and *high*. For experimentation, we test well-known machine learning algorithm on our dataset and compare the results to find which algorithm is the most suitable for our data and the problem scenario.

4.2 Experiments and Results

For this experiment we take 2 datasets. The first dataset consists of only the traditional attributes and we call it “Traditional”. The second dataset consists of traditional attributes along with social media attributes and we call it “Combined”. The description of the datasets with an example has been given in the tables below.

TABLE 3- Combined Dataset description with an example

TYPE	Film	Rotten Tomatoes	Audience Score	#Theaters Opening Week	Box Office Avg($)	Budget($M)	Worldwide Gross($M)	Facebook Likes	Youtube Likes	Twitter Followers	Class
Combined	Sanctum	28	48	3777	16213	125	493.21	14065 01	126	18382	Medium

TABLE 4- Traditional dataset description with an example

TYPE	Film	Rotten Tomatoes	Audience Score	#Theaters Opening Week	Box Office Avg($)	Budget($M)	Worldwide Gross($M)	Class
Traditional	Sanctum	28	48	3777	16213	125	493.21	Medium

We used RapidMiner version5.0 and Weka v3.6 for experimentation purpose. In this section, we shall explain the experimental setup for the algorithms we tested on our dataset one by one under the headings of Process Flow Diagrams, Settings, Parameters and Results.

DECISION TREE

Algorithm: Decision Tree

Settings: 10 fold Cross-Validation with Shuffled sampling

Parameters: Splitting Criterion: Gini-Index, Minimal Gain: 0.001

Process Flow:

FIGURE 6- Cross-validation step for Decision Tree

Figure 6 [45] above describes the process flow of the experiment. Here, we see data is input for Cross-Validation with the settings mentioned above. It outputs the output model, training set meta-data view, performance vector and result overview. We also note a double-square in the right corner of the 'validation box'. This sign denotes it is a nested operator, which means it has two sub-processes: a training sub-process and a testing sub-process inside itself.

FIGURE 7- Training and Testing Phase for Decision Tree

In Figure 7 [46], the training sub-process is used for training a model. The trained model is then applied in the testing sub-process. Thus, we input the training set into Decision Tree model with the settings mentioned above. The training model is output from the training process into the testing process where the model is applied on the test set through 'Apply Model' box. Hence, the performance of the model is also measured during the testing phase through 'Performance' box.

Decision Tree learner: A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a *rooted tree* with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an *internal* or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values[41]. We use Gini index which is an

impurity-based criterion that measures the divergences between the probability distributions of the target attribute's values. Using Gini-Index as the splitting criterion, we choose attributes at each stage which defines the construction of the tree.

Result:

TABLE 5- Comparing Classification Accuracy (F-Score) for Traditional and combined Datasets using Decision Tree (N=532)

	Low	**Medium**	**High**	**Average**
Traditional	0.78	0.39	0.27	0.48
Combined	0.91	0.68	0.64	0.74

In order to boost the performance, we use some ensemble techniques. Below are the details on the implementation, settings and results.

RANDOM FOREST

Algorithm: Random Forest

Settings: 10 fold Cross-Validation with Shuffled sampling, No. of trees: 10

Parameters: Splitting Criterion: Gini-Index, Minimal Gain: 0.001

Process Flow:

FIGURE 8- Cross-validation step for Random Forest

Figure 8[45] describes the cross-validation step for Random Forest algorithm with details being the same as for Figure 6 discussed above

FIGURE 9- Testing and Training phase for Random Forest

Here in Figure 9 [47], we input the training set into Random Forest model with the settings mentioned above. Rest of the settings and process flow for training and testing phase is the same as described for Figure 7.

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a many decision trees for training and outputting the class that is the mode of the classes output by individual trees. They use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. [42].

Result:

TABLE 6- Comparing Classification Accuracy (F-Score) for Traditional and combined Datasets using Random Forest (N=532)

	Low	**Medium**	**High**	**Average**
Traditional	0.81	0.41	0.34	0.60
Combined	0.84	0.62	0.57	0.756

The average F-Score for Random Forest is the weighted average where classes have been assigned weight according to their size. Next, making an attempt to improve the performance of Decision Tree, we tried the ensemble technique of 'bagging'. In the following experiment, this technique has been illustrated in details.

BAGGED DECISION TREE

Algorithm: Bagged Decision Tree

Settings: 10 fold Cross-Validation with Shuffled sampling

Parameters:

Bagging: Sampling Ratio : 0.9, Iterations: 10

Decision Tree learner: Splitting Criterion- Gini-Index, Minimal Gain: 0.001

Process Flow:

FIGURE 10- Cross-Validation phase for Bagged Decision Tree

Figure 10[45] describes the cross-validation step for Bagged Decision Tree algorithm with details being the same as for Figure 6 discussed above.

FIGURE 11- Training and Testing phase using Decision Tree learner inside 'Bagged' meta-modeling operator

Here in Figure 11 [48], we input the training set into the 'Bagged' meta-modeling operator with the settings mentioned above. Next, 'Decision Tree' learner is fed into the meta-modeling operator. Rest of the settings and process flow for training and testing phase is the same as described for Figure 7.

Bagging is a special case of the model averaging approach. **Bootstrap aggregating** (**bagging**) is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid over-fitting. Given a standard training set *D* of size *n*, bagging generates *m* new training sets D_i, each of size *n'*, by sampling from *D* uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i. If *n*'=*n*, then for large *n* the set D_i is expected to have the fraction (1 - 1/*e*) (≈63.2%) of the unique examples of *D*, the rest being duplicates. This kind of sample is known as a bootstrap sample. The *m* models

are fitted using the above *m* bootstrap samples and combined by averaging the output (for regression) or voting (for classification)[41].

Result:

TABLE 7- Comparing Classification Accuracy (F-Score) for Traditional and combined Datasets using Bagged Decision Tree (N=532)

	Low	**Medium**	**High**	**Average**
Traditional	0.81	0.40	0.24	0.483
Combined	0.92	0.72	0.613	0.75

ADAPTIVE BOOSTED DECISION TREE

Algorithm: Adaptive Boosted Decision Tree

Settings: 10 fold Cross-Validation with Shuffled sampling

Parameters:

Boosting: Iterations: 10

Decision Tree learner: Splitting Criterion- Gini-Index, Minimal Gain: 0.001

Process Flow:

FIGURE 12- Cross-Validation phase for Adaptive Boosted Decision Tree

Figure 12[45] describes the cross-validation step for Adaptive Boosted Decision Tree algorithm with details being the same as for Figure 6 discussed above.

Figure 13- Training and Testing phase using Decision Tree learner inside ‘Adaptive Boost’ meta-modeling operator

Here in Figure 13 [48], we input the training set into the ‘Boosted’ meta-modeling operator with the settings mentioned above. Next, ‘Decision Tree’ learner is fed into the

meta-modeling operator. Rest of the settings and process flow for training and testing phase is the same as described for Figure 7.

Boosting is a machine learning meta-algorithm for reducing bias in supervised learning. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that becomes the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. It is sensitive to noisy data and outliers. Also, it is very popular and perhaps the most significant as it was the first algorithm that could adapt to the weak learners[43].

Result:

TABLE 8- Comparing Classification Accuracy (F-Score) for Traditional and combined Datasets using Adaptive Boosted Decision Tree (N=532)

	Low	**Medium**	**High**	**Average**
Traditional	0.803	0.39	0.18	0.457
Combined	0.90	0.63	0.51	0.68

SUPPORT VECTOR MACHINE

Algorithm: Support Vector Machine Learner using mySVM developed by Stefan Ruping

Settings: 10 fold Cross-Validation with Shuffled sampling

Parameters:

C (Penalty of Misclassification):5.0, Gamma:1.0.

Polynomial to Binomial Learner:1-vs-all strategy

Process Flow: Figure 13 describes the cross-validation step for the Support Vector Machine algorithm with details being the same as for Figure 11 discussed above

FIGURE 14- Cross-Validation Step with Support Vector Machine

FIGURE 15- Training and Testing phase using SVM learner inside 'Polynomial to Binomial' meta-modeling operator to implement multiclass SVM

The Polynomial by Binomial Classification operator is a meta-modeling operator i.e. it has a sub-process. The sub-process must have a binomial classification learner. This operator builds a polynomial classification model using the binomial classification learner provided in its sub-process.

Here in figure 15 [49], we input the training set into 'Polynomial to Binomial' meta-modeling operator with the settings mentioned above. Next, 'Support Vector Machine' learner is fed into the meta-modeling operator Rest of the settings and process flow for training and testing phase is the same as described for Figure 7. The settings for Figure 14 [45] is the same as Figure 6 above.

Support Vector Machine: A SVM training algorithm is a non-probabilistic binary linear classifier where a model that assigns new examples into one category or the other, making it a. A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class. Multiclass SVM is building binary classifiers which distinguish between (i) one of the labels and the rest (*one-versus-all*) or (ii) between every pair of classes (*one-versus-one*). Classification of new instances for the one-versus-all case is done such that the classifier with the highest output function assigns the class. For the one-versus-one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes,

then the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the instance classification[44].

The (**Gaussian**) **radial basis function kernel**, or **RBF kernel**, is a popular kernel function used in support vector machine classification.[40] The RBF kernel on two samples **x** and **x'**, represented as feature vectors in some *input space*, is defined as

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||_2^2}{2\sigma^2}\right)$$

$||\mathbf{x} - \mathbf{x}'||_2^2$ may be recognized as the squared Euclidean distance between the two feature vectors. σ is a free parameter. An equivalent, but simpler, definition involves a parameter $\gamma = -\frac{1}{2\sigma^2}$:

$$K(\mathbf{x}, \mathbf{x}') = \exp(\gamma||\mathbf{x} - \mathbf{x}'||_2^2)$$

Result:

TABLE 9- Comparing Classification Accuracy (F-Score) for Traditional and combined Datasets using Multiclass SVM (N=532)

	Low	**Medium**	**High**	**Average**
Traditional	0.835	0.68	0.366	0.59
Combined	0.941	0.696	0.625	0.754

4.3 Performance Analysis

We analyzed the results we got with focus on data and algorithm separately. With focus on the data part, we use our general knowledge to reason out for better classification performance with social media opposed to with traditional variables.

Film	Rotten Tomatoes	Audience Score	#Theaters Opening Week	Box Office Avg($)	Budget($M)	Worldwide Gross($M)	Facebook Likes	Youtube Likes	Twitter Followers	Class
Harry Potter and Deadly Hallows Part1	79	87	4125	30307	125	955	126588 2	13439 5	163743 3	High
Sanctum	28	48	3777	16213	125	493.21	140650 1	126	18382	Medium
The Taking of Pelham 1 2 3	64	68	3611	15291	130	185.30	24798	4224	56002	Low
17 Again	55	70	3255	7288	20	136.30	356495 4	3499	379594 8	High
Bride Wars	11	56	3226	6528	30	115.15	926987	378	108816	Medium
She's Out of my League	57	60	2956	3307	20	48.81	317306	2263	44654	Low
The Hangover Part II	35	58	3615	23775	80	581.46	162023 4	3515	115808 4	High
Grown Ups	10	59	3534	11462	80	267.4	692731 4	2869	642167	Medium
Zookeeper	14	42	3482	5763	80	170.30	945954	1086	127227	Low

Table 10- A random collection of movies, which fall in Low, Medium and High profitability class

For a movie to become a blockbuster, it is definite that there are values of some factors other than the mainstream ones which are deterministic in influencing the box office performance of movies. Based on the above understanding, we investigated our dataset again to check for the variables which could explain the high profitability of the movie.

Upon careful examination, we found:

- Popularity index tapped through *Facebook likes*, *Youtube likes* and *Twitter followers* were independent of the budget or number of theaters the movie was released in.

- The data extracted from social media and multimedia sites are significantly influential in describing the profit made by the movie.

Hence, we conclude:

- The social signal does add some information about the movie's potential at the box office beyond traditional attributes.

- Complimentary to the findings of [11], we also suggest that variables such as budget, number of theaters on opening week, genre, studio etc do not play a significant role in driving the box office performance.

Thus our results indicate the strength of social media in influencing the box office revenue of movie, especially with movies which make huge success at the theaters.

With focus on the algorithms performance, we see that SVM outperformed all other algorithms we tested our dataset on. We started with Decision Tree using CART algorithm for classification purpose taking reference from some previous studies which suggest Decision Tree performs well for classification jobs. Though Decision Tree performed well with the dataset, we tried to apply some more techniques in order to improve its performance. We tested the performance of dataset with Random Forest and the performance became better by a small margin. Seeking for improvement, we next tried the bagging technique to improve performance and we see it leads to slightly

improve results. Many decision trees help develop a better model than an individual decision tree. Making an attempt to further improve on the performance we tried Adaptive Boosting on Decision Tree algorithm. Unfortunately, we see the performance of the model to have decreased. We think this could be due to noise in the data as it is real world data and so cannot expect it to be perfect. We next tested the dataset on Support vector Machine seeking for improved classification. Interestingly, SVM learner trained using Radial basis function suits our dataset. We see significant improvement in the accuracy with which the model performs which possibly is influenced by the nature of the dataset. RBF fits better for real world dataset which are linearly inseparable due to unequal class distribution and some overlaps between classes [38]. Also, RBF kernel has been studied to perform better with small number of features in the dataset and small training samples in input space.

CHAPTER 5

DETERMINE THE MOST RELEVANT SOCIAL SIGNAL INFLUENCING THE PERFORMANCE OF MOVIES

5.1 Introduction

In this chapter of this thesis, we shall begin with defining our motivation towards setting this goal, the contribution it shall make in understanding this problem scenario better, our methodology, followed by the experimental results with analysis and conclusion.

Previous research has touched upon locating social signals to tap popularity index of stars through polls or buzz, popularity of movies through Wikipedia activities, tweets, Youtube and Google searches etc. In spite of having collected a wide variety of social signals, we see that previous studies have somehow failed to pick signals which could reflect with reasonable accuracy the performance of low and medium popular movies. This can be understood due to lack of statistically significant values of the metrics/signals chosen to do analysis for such low/medium popular movies. In [8], the authors have used buzz in form of tweets one night before the release of the movie. It can be easily understood that low/medium popular movies lack the ability to create enough buzz on platforms like Twitter which makes the tweet signal skewed towards highly popular movies. Similar could be the situation with activity level on Wikipedia articles for movies where significant amount of edits are not made on low/medium popular movies owing to low interest of the people in the movie. We understand that there is some minimal amount of data needed to do statistical analysis of our kind and thus, in our study, we aim at choosing such signals which are accessible to the largest population of people, and so,

could provide statistically significant numbers for low/medium popular movies as well. We have collected signals from social media which reflect popularity of movies and stars in different perspectives. We also do another comparison study to test which social media signal is the strongest predictor of movie profitability. We thought this was an interesting problem to study, given that previous papers have not explored the subject in the light of the particular social media signals we use.

Motivation

Our motivation towards conducting this study was to help the Motion Picture Industry understand the most significant social signal which could predict the performance of movies.

Contribution

Our study will not only help studios and directors determine which components related to movies could be the most crucial in influencing its box-office success, but also, focus on that social signal our study reveals for better promotion and advertisement.

Methodology

We test some well-known machine learning algorithms on our dataset as we did earlier. Based on our results, we analyze and determine the social signal which is the most relevant to the performance of the movies, along with the algorithm which gives the best results.

5.2 Experiments and Results:

The setup for the experiments remains the same as in Chapter 4 except for the dataset which is different. The tables below describe how the datasets look like with the help of an example.

TABLE 11- Traditional+Twitter Dataset description with an example

TYPE	Film	Rotten Tomatoes	Audience Score	#Theaters Opening Week	Box Office Avg($)	Budget($M)	Worldwide Gross($M)	# of Twitter Followers	Class
Traditional + Twitter	Sanctum	28	48	3777	16213	125	493.21	18382	Medium

TABLE 12- Traditional + Facebook Dataset description with an example

TYPE	Film	Rotten Tomatoes	Audience Score	#Theaters Opening Week	Box Office Avg($)	Budget($M)	Worldwide Gross($M)	# of Facebook Likes	Class
Traditional + Facebook	Sanctum	28	48	3777	16213	125	493.21	1406501	Medium

TABLE 13- Traditional + Youtube Dataset description with an example

TYPE	Film	Rotten Tomatoes	Audience Score	#Theaters Opening Week	Box Office Avg($)	Budget($ M)	Worldwide Gross($M)	# of Youtube Likes	Class
Traditional + Youtube	Sanctum	28	48	3777	16213	125	493.21	126	Medium

We use the same pattern as in Chapter 4.2 to illustrate our experiments and results.

Decision Tree

TABLE 14- Comparing Classification Accuracy (F-Score) for Traditional variables along with the different social variables taken individually using Decision Tree (N=532)

Dataset	Low	Medium	High	Average
Traditional+Facebook	0.906	0.704	0.605	0.738
Traditional+Twitter	0.907	0.69	0.67	0.75
Traditional+YouTube	0.919	0.685	0.65	0.74

Random Forest

TABLE 14- Comparing Classification Accuracy (F-Score) for Traditional variables along with the different social variables taken individually using Random Forest (N=532)

Dataset	Low	Medium	High	Average
Traditional+Facebook	0.90	0.63	0.49	0.74
Traditional+Twitter	0.914	0.65	0.51	0.79
Traditional+YouTube	0.901	0.62	0.482	0.76

Bagged Decision Tree

TABLE 15- Comparing Classification Accuracy (F-Score) for Traditional variables along with the different social variables taken individually using Bagged Decision Tree (N=532)

Dataset	Low	Medium	High	Average
Traditional+Facebook	0.92	0.67	0.50	0.696
Traditional+Twitter	0.93	0.76	0.73	0.806
Traditional+YouTube	0.93	0.75	0.66	0.78

Adaptive Boosted Decision Tree

TABLE 16- Comparing Classification Accuracy (F-Score) for Traditional variables along with the different social variables taken individually using Adaptive Boosted Decision Tree (N=532)

Dataset	Low	Medium	High	Average
Traditional+Facebook	0.91	0.66	0.46	0.67
Traditional+Twitter	0.91	0.73	0.60	0.746
Traditional+YouTube	0.91	0.678	0.47	0.69

Support Vector Machine

TABLE 17- Comparing Classification Accuracy (F-Score) for Traditional variables along with the different social variables taken individually using Multiclass SVM (N=532)

Dataset	Low	Medium	High	Average
Traditional+Facebook	0.90	0.75	0.67	0.77
Traditional+Twitter	0.975	0.886	0.82	0.893
Traditional+YouTube	0.94	0.76	0.72	0.80

5.3 Performance Analysis

The average F-Score values when taking social signals one at a time are consistently higher than that for “Combined” dataset where all three social signals are taken together

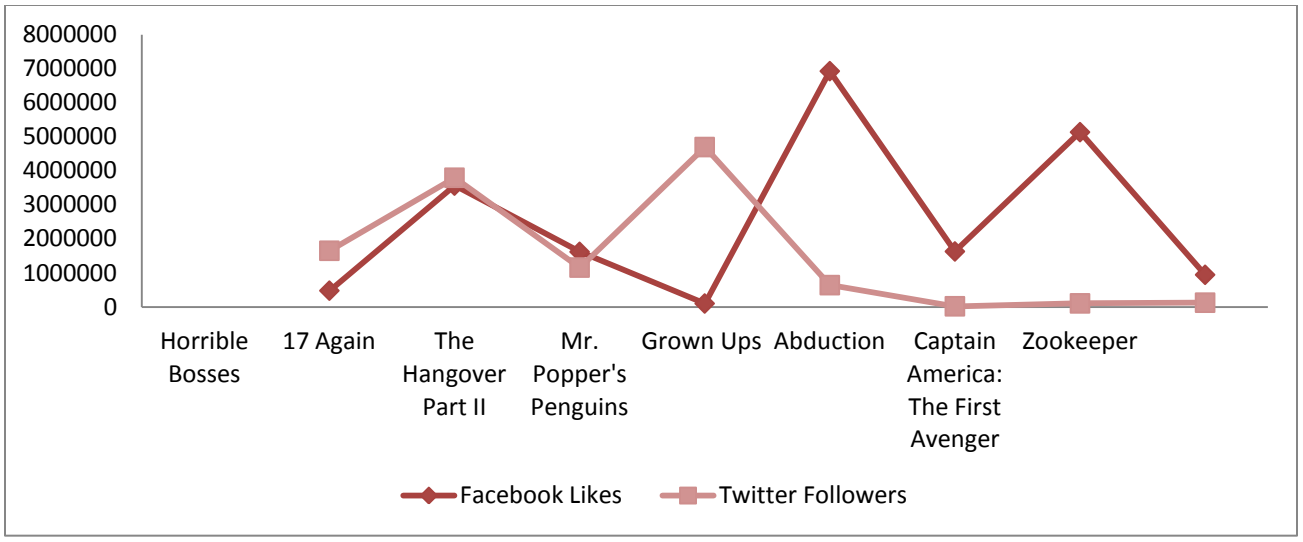

FIGURE 16- Facebook like and Twitter follower count for movies shown

A plausible reason for such result could be amplified noise caused when the signals are used together as we see in Figure 16 a high degree of disassociation between Facebook like and Twitter follower signal for a random set of movies chosen representing different profitability classes. This weakens the quality of the model generated.

We outline the following conclusion out of the experiments.

a. *Popularity of actors, tapped through follower count on Twitter is a significant factor, which drives the performance of movie.*

Our result thus helps shed light on the long-standing debate about star power in determining box-office success of movies. A number of previous works have used traditional methods such as magazine polls, Harris polls etc. Others have used tweets related to movies or stars as proxy for actors' quality or popularity. Our result is indicative that the strongest social signal to influence the performance of movie is the *popularity of actors captured through follower count of movie casts on Twitter*. Upon

comparison, we see that Twitter follower count for actors of the movie help classify *movies with overall F-Score being 0.89 and in high profitability class with 0.82.*

Our result shows feature engineering is important, since the follower count of movie cast is a better feature than individual tweets about movie. Tweets about movies suffer inconsistent hype-approval factor which cannot be necessarily correlated with the financial performance of movies at box-office [7]. We also improve in terms of the accuracy of classification of movies.

b. *Facebook 'like' signal has noise that interferes with proper classification.*

Low F-Score values with Facebook *'like'* signal, despite having the largest number of users amongst all social media looks strange and unexpected. A possible reason for this result could be that movie pages get likes from people of countries where the movie does not release in theaters and so those likes do not really correspond to ticket sales for the movie. Moreover, now-a-days Facebook pages for movies are made 9-10 months before the release of the movies. The page keeps garnering likes over-time which do not necessarily turn into ticket sale upon release.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Our work shows that social media influences the performance of movies significantly and adds more information about profitability of movies. In this thesis, we do not propose a new machine learning algorithm, rather we use carefully chosen social media data and evaluate the performance of well-known algorithms in light of the data. We show that popularity of actor depicted through follower count on Twitter is most relevant to the success of movie at theaters. Another very interesting result we have is that Facebook likes cannot be counted to be a credible signal for similar analysis. Our results suggest that the *like* signal has noise which impedes its analytical capability. Finally, we see that modern classification algorithm Support Vector Machine performs better than other classification algorithms.

This study sets ground for future researchers to further investigate and potentially exploit other facets of the fact that follower count on Twitter is a credible measure of popularity. It could be interesting to assess popularity through follower count of other aspects involved in the film like production studio or director of the movie etc. Also, our result indicates that Twitter can be a great platform to measure popularity. We encourage future researchers to tap the fan/follower count for movies through their welcome page on Twitter. With rigorous research going on in developing classification algorithms to determine bots, fake accounts and humans on Twitter, we expect the data in the future to be cleaner and more credible. Also, though Facebook likes could not prove to be a strong indicator to assess financial performance of movies, we also believe that the

recommendation power of Facebook can be further investigated by using its OpenGraph Protocol.

Another interesting target to be achieved with this work could be to predict the profitability class of the movie which are scheduled to be released by gathering the data pertaining to them and checking the change in the values of social signal as the movie approaches the release date. This might help understand the rise in the movie's and star's popularity with the release of movie coming nearer.

REFERENCES

[1] Eliashberg, J., Hui, S.K. and Zhang, Z.J., "From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts," *Management Science*, vol. 53, no. 6, pp. 881–893, 2007.

[2] Litman, B.R., "Predicting Success of Theatrical Movies: An Empirical Study", *Journal of Popular Culture*, vol. 16, no. 9, pp. 159-175, 1983.

[3] Ravid, S.A., "Information, Blockbusters, and Stars: A Study of the Film Industry", *Journal of Business*, vol. 72, no. 4, pp. 463-492, 1999.

[4] Mestyán, M., Yasseri, T., and Kertész, J.,"Early Prediction of Movie Box Office Success based on Wikipedia Activity Big Data", arXiv preprint arXiv:1211.0970, 2012.

[5] Reggie Panaligan, Andrea Chen (June 2012). "Quantifying Movie Magic with Google Search" [Online].

[6] Joshi M, Das D, Gimpel K, Smith N,"Movie reviews and revenues: An experiment in text regression", in *Proceedings of NAACL-HLT*, PA, pp. 293-296, 2010.

[7] Wong FMF, Sen S, Chiang M.,"Why watching movie tweets won't tell the whole story?" *in Proceedings of the 2012 ACM workshop on Workshop on online social networks*, NY,USA, pp. 61-66, 2012.

[8] Asur S, Huberman BA,"Predicting the future with social media" in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, DC, pp. 492-499, 2010.

[9] Karniouchina, Ekaterina, "Impact of star and movie buzz on motion picture distribution and box office revenue*", Intern. J. of Research in Marketing*, vol. 28, pp. 62-74, 2011

[10] Spann, Martin & Bernd Skiera," InternetBased Virtual Stock Markets for Business Forecasting", *Management Science* , vol. 49, no. 10, pp. 1310-1326, 2003.

[11] Kaplan, Joshua J., "Turning Followers into Dollars: The Impact of Social Media on a Movie's Financial Performance", *Undergraduate Economic Review*, vol. 9, no. 1, Article 10, 2012.

[12] O'Dell, Jolie "Twitter: The Killer Box Office Predictor*?"*, pp. 553 – 556, 2010. Available http://mashable.com/2010/04/02/twitter-the-killer-box- office-predictor-2/

[13] Huang, Jin et. al., "Comparing Naïve Bayes, Decision trees and SVM with AUC and accuracy. Data Mining", *Third IEEE International Conference on Data Mining*, pp. 553-556, Nov 2003.

[14] Amit Sharma et.al., "Do Social Explanations Work? Studying and Modeling the Effects of Social Explanations in Recommender Systems", in *Proceedings of the 22nd international conference on World Wide Web*, Geneva, Switzerland, pp. 1133-1144, 2013.

[15] How to recognize Twitter bots: 7 signals to look out for, http://www.stateofdigital.com/how-to-recognize-twitter-bots-6-signals-to-look-out-for/

[16] Shugan, Steven M. & Joffre Swait (2000). Enabling Movie Design and Cumulative Box Office Predictions. ARF Conference Proceedings.

[17] Eliashberg, Jehoshua (2004). The Film Exhibition Business: Critical Issues, Practice and Research. In Charles C. Moul (Editor). A Short Handbook of Movie Industry Economics. New York City, New York: Cambridge University Press.
[18] Krider, R. E., & Weinberg, C. B. (1998). Competitive Dynamics and the Introduction of New Products: The Motion Picture Timing Game. Journal of Marketing Research, 35 (February), 1-15.
[19] Chisholm, D. C. (2000). The War of Attrition and Optimal Timing of Motion-Picture Releases. Working Paper, Lehigh University, (July 2000).
[20] Sochay, S. (1994). Predicting the Performance of Motion Pictures. Journal of Media Economics, 7(4), 1-20.
[21] Smith, S. P., & Smith, V. K. (1986). Successful Movies: A Preliminary Empirical Analysis. Applied Economics, 18, 501-507.
[22] Gruca, Thomas (2000). The IEM Movie Box Office Market Integrating Marketing and Finance using Electronic Markets. Journal of Marketing Education, 22: 5-14.
[23] Elberse, Anita & Eliashberg, Jehoshua (2003). Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures. Marketing Science 22 (3, Summer), 329-354.
[24] Sood, Sanjay and Xavier Dreze (2004). Brand Extensions of Hedonic Goods: Movie Sequel Evaluations. Working Paper.
[25] Litman, B. R., & Kohl, L. S. (1989). Predicting Financial Success of Motion Pictures: The '80s Experience. Journal of Media Economics, 2, 35-50.
[26] Litman, B. R., & Ahn, H. (1998). Predicting Financial Success of Motion Pictures. B. R. Litman The Motion Picture Mega-Industry. Needham Heights, MA: Allyn & Bacon.
[27] Wallace, W. T., Seigerman, A., & Holbrook, M. B. (1993). The Role of Actors and Actresses in the Success of Films: How Much is a Movie Star Worth? Journal of Cultural Economics, 17(1), 1-27.
[28] Albert, Steven (1998). Movie Stars and the Distribution of Financially Successful Films in the Motion Picture Industry. Journal of Cultural Economics, 22, 249-270.
[29] De Vany, A., & Walls, W. D. (1996). Bose-Einstein Dynamics and Adaptive Contracting in the Motion Picture Industry. The Economic Journal, 106(November), 1493-1514.
[30] Brewer, Stephanie & Jozefowicz, James & Kelley, Jason (2009). *A blueprint for success in the US film industry*. Applied Economics, *41*, 589-606.
[31] S. D. Roy, T. Mei, W. Zeng and S. Li., "Empowering Cross-Domain Internet Media With Real-Time Topic Learning From Social Streams." Proceedings of *IEEE International Conference on Multimedia and Expo*, (2012).
[32] Achrekar, Harshavardhan, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. "Predicting flu trends using twitter data." In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pp. 702-707. IEEE, (2011) .
[33] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." *Proceedings of the 19th international conference on World wide web*. ACM, (2010).
[34] Biever, Celeste. "Twitter mood maps reveal emotional states of America." *New Scientist* 207.2771 (2010): 14.

[35] B. A. Huberman, D. A. Romero and F. Wu. 2008. "Social networks that matter: Twitter under the microscope," In *Computing Research Repository - CORR*, vol. abs/0812.1, no. 1, (2008)

[36] Ishii, Akira, Hisashi Arakaki, Naoya Matsuda, Sanae Umemura, Tamiko Urushidani, Naoya Yamagata, and Narihiko Yoshida, "The 'hit' phenomenon: a mathematical model of human dynamics interactions as a stochastic process," New Journal of Physics 14.6 (2012): 063018

[37] Mishne G, Glance N (2006) Predicting movie sales from Blogger sentiment. In: Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAICAAW).

[38] Daehyon KIM et. al., "Performance Improvement in Traffic Vision Systems using", Journal of the Eastern Asia Society for Transportation Studies, Vol. 6, pp. 2589 - 2599, 2005

[39] Suman D. Roy and W. Zeng, "The Hidden Potential of Movie Genome Communities: Analyzing fine-grained Semantic Information in Motion Pictures," in *IEEE International Conference on Semantic Computing*, Irvine, CA, Sept. 2013.

[40] Wikipedia, http://en.wikipedia.org/wiki/RBF_kernel

[41] Maimon, Oded Z., and Lior Rokach, eds. Data mining and knowledge discovery handbook. Vol. 1, Chapter 9, pg.149, New York: Springer, 2005.

[42] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[43] Schapire, Robert E. "The boosting approach to machine learning: An overview."Nonlinear estimation and classification. Springer New York, 2003. 149-171.

[44] Maimon, Oded Z., and Lior Rokach, eds. Data mining and knowledge discovery handbook. Vol. 1, Chapter 12, pg.232, New York: Springer, 2005.

[45] Rapidminer Operator Reference, http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/RapidMiner_OperatorReference_en.pdf, pg. 896

[46] Rapidminer Operator Reference, http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/RapidMiner_OperatorReference_en.pdf, pg. 659

[47] Rapidminer Operator Reference, http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/RapidMiner_OperatorReference_en.pdf, pg. 680

[48] Rapidminer Operator Reference, http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/RapidMiner_OperatorReference_en.pdf, pg. 754,757

[49] Rapidminer Operator Reference, http://1xltkxylmzx3z8gd647akcdvov.wpengine.netdna-cdn.com/wp-content/uploads/2013/10/RapidMiner_OperatorReference_en.pdf, pg. 714