# Evaluation of Actors, Supporting Actors, and Directors' Comprehensive profiles from Social Media and Awards to Predict Hollywood Movies Success

*By*

**Sumbal Khan**

MCS153017

**MASTER OF SCIENCE IN COMPUTER SCIENCE**

**DEPARTMENT OF COMPUTER SCIENCE**
**CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY**
**ISLAMABAD**
**2017**

# Evaluation of Actors, Supporting Actors, and Directors' Comprehensive profiles from Social Media and Awards to Predict Hollywood Movies Success

*By*

**Sumbal Khan**

A research thesis submitted to the
Department of Computer Science, Capital University of Science and Technology, in partial fulfillment of the requirements for the degree of

**MASTER OF SCIENCE IN COMPUTER SCIENCE**

**DEPARTMENT OF COMPUTER SCIENCE**
**CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY**
**ISLAMABAD**
**2017**

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD

## CERTIFICATE OF APPROVAL

**Evaluation of Actors, Supporting Actors, and Directors' Comprehensive profiles from Social Media and Awards to Predict Hollywood Movies Success**

By

Sumbal Khan

Reg No. MCS153017

**THESIS EXAMINING COMMITTEE**

| S No | Examiner | Name | Organization |
|---|---|---|---|
| (a) | External Examiner | Dr. Majid Iqbal Khan | CIIT, Islamabad |
| (b) | Internal Examiner | Dr. Nayyer Masood | CUST, Islamabad |
| (c) | Supervisor | Dr. Muhammad Tanvir Afzal | CUST, Islamabad |

Dr. Muhammad Tanvir Afzal

**Thesis Supervisors**

October, 2017

Dr. Nayyer Masood
Head
Department of Computer Science
Dated : October, 2017

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
Dated : October, 2017

# Certificate

This is to certify that Sumbal Khan has incorporated all observations, suggestions, and comments made by the external as well as the internal examiners and thesis supervisors. The title of her thesis is: **Evaluation of Actors, Supporting Actors, and Directors' Comprehensive profiles from Social Media and Awards to Predict Hollywood Movies Success**

Forwarded for necessary action

Dr. Muhammad Tanvir Afzal

(Thesis Supervisor)

*I dedicate my dissertation work to my family, teachers and friends. Special feeling of gratitude to my loving parents for their love, endless support and encouragement.*

**Acknowledgement**

All worship and praise is for ALLAH (S.W.T), the creator of whole worlds. First of all, I would like to thank Him for countless blessings and for providing me the strength and knowledge to complete my research. Secondly, special thanks to my supervisor Professor **Dr. Muhammad Tanvir Afzal** for his excellent supervision and great support during my study. Not only teaching me how to conduct research, he has also given me an insight into my future career. I have enjoyed a lot doing research because of my supervisor and a highly motivating research topic.

I would like to thank my family for their love and moral support, and especially my brothers and sister for inspiration and always being there when I needed.

I would like to thank **Mr. Usman Ahmed** also for his help and valuable input during the conduct of my research.

Finally, I would also like to thank my roommates **Mahak Manzoor** and **Rimsha Butool** to help me and keep me motivated during this study. Although I have been through some difficult time during my research, however they always insist not to give up and help me completing this research Thesis.

**Sumbal Khan**

**MCS153017**

# DECLARATION

I, hereby declare that this thesis neither as a whole nor as a part thereof has been copied out from any source. It is further declared that I have developed this research and the accompanied report entirely on the basis of my personal efforts made under the sincere guidance of my supervisor. No portion of the work presented in this report has been submitted in the support of any other degree or qualification of this or any other University or Institute of learning, if found we shall stand responsible.

Sumbal Khan

Reg. No, MCS153017

October, 2017.

Abstract

The Hollywood film industry releases number of movies every year. However, only a few movies taste success and ranked high. Production of a successful movie is not an easy task. Hollywood industry has to release such successful movies which are quite entertaining for the audience. The questions arise that how to predict that a particular movie will be entertaining for the audience and is there any way to predict the success of a movie before its release or even before its production. Large amount of data related to the movies is available over the internet, because it is an interesting data mining topic nowadays. Data Analysts and movie maker constantly feel a need to have an expert system, which can forecast the movie success with reasonable accuracy.

Movie success prediction has extensively studied by the experts which include data analyst, econometricians as well as a marketing professional. Generally, the variable used in their research includes production budgets, pre-release advertising expenditures, run time, and seasonality. By using different approaches, these forecasting models forecast the financial success of movies, but most of them are targeted post-production phase or have low prediction F-measure. These variables are time dependent and are only getable when story, director and cast are finalized. This reason is that when model forecast the success of movie, investor money already been spend and didn’t carry any meaningful impact. These models have a limited scope and non-ability to reduce revenue loss risk.

Pre-release prediction is only possible when we have historical data. The movie cast performance can be evaluated by seeing the number of awards won by the lead cast member. A number of studies have related to movies using social networking (Twitter) however; less significant work has been done using Academy Awards. No researches relating to movie prediction using Instagram and other awards such as Golden Globe and Venice Awards have been evaluated. The focal point of this research is to analyze that how different award Oscar, Golden Globe and Venice Awards and cast social (twitter, Facebook & Instagram) media popularity can be used to predict success of the movie. These feature prediction powers are taken into account for while selection.

In this study, Hollywood movies data for last 10 years (2005-2015) were collected. Data were collected from IMDB, Facebook, Instagram and Twitter. After pre-processing, two categories were evaluated; the first category discusses the F-measure of features in the dataset using output classes (A-I) and the second category discuss the F-measure with only 2 classes A and B. All of the features have been evaluated as independent features and in combinations as well. Four classifiers have been used in this thesis such as: Random Forest, Naïve Bayes, Support Vector Machine and Decision Tree (J48).

# Contents


LIST OF FIGURES ..... xiii
LIST OF TABLES ..... xv
LIST OF ACRONYMS ..... xvi
Chapter 1: Introduction ..... 1
1.1 Purpose ..... 3
1.2 Problem Statement ..... 3
1.3 Scope ..... 4
1.4 Significance of the Solution ..... 4
1.5 Dissertation Organization ..... 4
1.6 Definitions and Abbreviations ..... 4
1.7 Conclusion ..... 5
Chapter 2: Literature Review ..... 6
2.1 Social Media ..... 8
2.1.1 Facebook ..... 8
2.1.2 Twitter ..... 9
2.1.3 Instagram ..... 9
2.2 Prestigious Award ..... 9
2.3 Social Media Marketing Strategy ..... 9
2.4 Forecasting of Motion Picture Revenue ..... 15
2.4.1 Post-Production Forecasting Models ..... 16
2.4.2 Pre-Production Forecasting Models ..... 20
Chapter 3: Methodology ..... 22
3.1 Data Collection ..... 22
3.2 Feature Set ..... 25
3.3 Pre-Processing ..... 26
3.4 Experiments and Evaluation ..... 27
Chapter 4: Results ..... 29
4.1 Data Statistics ..... 29
4.2 Pre –Processing ..... 29
4.3 Evaluation ..... 30
4.4 Social Media Impact on Movie Success ..... 30
4.4.1 One Feature Analysis ..... 31
4.4.2 Two Feature Analyses ..... 39
4.4.3 Three Feature Analyses ..... 48

4.5 Awards Impact on Movie Success ........ 56
4.5.1 One Feature Analysis ........ 56
4.5.2 Two Feature Analyses ........ 65
4.5.3 Three Feature Analyses ........ 75
4.5.4 Conclusion ........ 83
Chapter 5: Conclusion and Future Work ........ 85
5.1 Conclusion ........ 85
5.2 Future Work ........ 87
References ........ 88

# LIST OF FIGURES


FIGURE 3.1 METHODOLOGY AND FEATURE SET ........ 24
FIGURE 3.2 MISSING VALUES ........ 27
FIGURE 4.1 DIRECTOR ONE FEATURE ANALYSIS FULL CLASSES ........ 32
FIGURE 4.2 DIRECTOR ONE FEATURE ANALYSIS TWO CLASSES ........ 33
FIGURE 4.3 ACTOR ONE- ONE FEATURE ANALYSIS FULL CLASSES ........ 34
FIGURE 4.4 ACTOR ONE- ONE FEATURE ANALYSIS TWO CLASSES ........ 35
FIGURE 4.5 ACTOR 2- ONE FEATURE ANALYSIS FULL CLASSES ........ 36
FIGURE 4.6 ACTOR 2- ONE FEATURE ANALYSIS TWO CLASSES ........ 37
FIGURE 4.7: ACTOR 3- ONE FEATURE ANALYSIS FULL CLASSES ........ 38
FIGURE 4.8: ACTOR 3- ONE FEATURE ANALYSIS TWO CLASSES ........ 39
FIGURE 4.9 DIRECTOR TWO FEATURE ANALYSES WITH TWO CLASSES ........ 40
FIGURE 4.10 DIRECTOR TWO FEATURE ANALYSES WITH ALL CLASSES ........ 41
FIGURE 4.11 ACTOR ONE- TWO FEATURE ANALYSIS WITH TWO CLASSES ........ 42
FIGURE 4.12 ACTOR ONE- TWO FEATURE ANALYSIS WITH ALL CLASSES ........ 43
FIGURE 4.13 ACTOR 2- TWO FEATURE ANALYSIS WITH TWO CLASSES ........ 45
FIGURE 4.14 ACTOR 2- TWO FEATURE ANALYSIS WITH ALL CLASSES ........ 46
FIGURE 4.15 ACTOR 3- TWO FEATURE ANALYSIS WITH TWO CLASSES ........ 47
FIGURE 4.16 ACTOR 3- TWO FEATURE ANALYSIS WITH ALL CLASSES ........ 48
FIGURE 4.17 DIRECTOR THREE FEATURE ANALYSES WITH TWO CLASSES ........ 49
FIGURE 4.18 DIRECTOR THREE FEATURE ANALYSIS WITH ALL CLASSES ........ 50
FIGURE 4.19 ACTOR ONE- THREE FEATURE ANALYSIS WITH TWO CLASSES ........ 51
FIGURE 4.20 ACTOR ONE- THREE FEATURE ANALYSIS WITH ALL CLASSES ........ 52
FIGURE 4.21 ACTOR 2- THREE FEATURE ANALYSIS WITH TWO CLASSES ........ 53
FIGURE 4.22 ACTOR 2- THREE FEATURE ANALYSIS WITH ALL CLASSES ........ 53
FIGURE 4.23 ACTOR 3- THREE FEATURE ANALYSIS WITH TWO CLASSES ........ 55
FIGURE 4.24 ACTOR 3- THREE FEATURE ANALYSIS WITH ALL CLASSES ........ 55
FIGURE 4.25 DIRECTOR ONE FEATURE ANALYSIS FULL CLASSES ........ 58
FIGURE 4.26 DIRECTOR ONE FEATURE ANALYSIS TWO CLASSES ........ 58
FIGURE 4.27 ACTOR ONE - ONE FEATURE ANALYSIS FULL CLASSES ........ 60
FIGURE 4.28 ACTOR ONE - ONE FEATURE ANALYSIS TWO CLASSES ........ 61
FIGURE 4.29 ACTOR 2 - ONE FEATURE ANALYSIS FULL CLASSES ........ 62
FIGURE 4.30 ACTOR 2 - ONE FEATURE ANALYSIS TWO CLASSES ........ 63
FIGURE 4.31 ACTOR 3 - ONE FEATURE ANALYSIS FULL CLASSES ........ 64
FIGURE 4.32 ACTOR 3 - ONE FEATURE ANALYSIS TWO CLASSES ........ 65
FIGURE 4.33 DIRECTOR TWO FEATURES ANALYSIS TWO CLASSES ........ 67
FIGURE 4.34 DIRECTOR TWO FEATURE ANALYSIS ALL CLASSES ........ 68
FIGURE 4.35 ACTOR ONE- TWO FEATURE ANALYSIS TWO CLASSES ........ 69
FIGURE 4.36 ACTOR ONE- TWO FEATURE ANALYSIS FULL CLASSES ........ 70
FIGURE 4.37 ACTOR 2- TWO FEATURE ANALYSIS TWO CLASSES ........ 72
FIGURE 4.38 ACTOR 2- TWO FEATURE ANALYSIS FULL CLASSES ........ 73
FIGURE 4.39 ACTOR 3- TWO FEATURE ANALYSIS TWO CLASSES ........ 74
FIGURE 4.40 ACTOR 3- TWO FEATURE ANALYSIS FULL CLASSES ........ 75

FIGURE 4.41 DIRECTOR THREE FEATURE ANALYSIS WITH TWO CLASSES .......... 76
FIGURE 4.42 DIRECTOR THREE FEATURE ANALYSIS WITH ALL CLASSES .......... 77
FIGURE 4.43 ACTOR ONE- THREE FEATURE ANALYSIS WITH TWO CLASSES .......... 78
FIGURE 4.44 ACTOR ONE- THREE FEATURE ANALYSIS WITH ALL CLASSES .......... 79
FIGURE 4.45 ACTOR 2- THREE FEATURE ANALYSIS WITH TWO CLASSES .......... 80
FIGURE 4.46 ACTOR 2- THREE FEATURE ANALYSIS WITH ALL CLASSES .......... 81
FIGURE 4.47 ACTOR 3- THREE FEATURE ANALYSIS WITH TWO CLASSES .......... 82
FIGURE 4.48 ACTOR 3- THREE FEATURE ANALYSIS WITH ALL CLASSES .......... 83

# LIST OF TABLES

TABLE 3.1 OUTPUT CLASS REVENUE RANGES ........ 23
TABLE 3.2 FEATURE SET ........ 26
TABLE 4.1 ONE FEATURE ANALYSIS DIRECTOR SOCIAL MEDIA ........ 31
TABLE 4.2: ONE FEATURE ANALYSIS LEAD ACTOR SOCIAL MEDIA ........ 33
TABLE 4.3: ONE FEATURE ANALYSIS SUPPORTING ACTOR 2 SOCIAL MEDIA ........ 35
TABLE 4.4 TWO FEATURE ANALYSIS DIRECTOR SOCIAL MEDIA ........ 39
TABLE 4.5: TWO FEATURE ANALYSIS LEAD ACTOR SOCIAL MEDIA ........ 41
TABLE 4.6: TWO FEATURE ANALYSIS SUPPORTING ACTOR 2 SOCIAL MEDIA ........ 44
TABLE 4.7: TWO FEATURE ANALYSIS SUPPORTING ACTOR 3 SOCIAL MEDIA ........ 46
TABLE 4.8: THREE FEATURE ANALYSIS DIRECTOR SOCIAL MEDIA ........ 48
TABLE 4.9: THREE FEATURE ANALYSIS LEAD ACTOR SOCIAL MEDIA ........ 50
TABLE 4.10: THREE FEATURE ANALYSIS SUPPORTING ACTOR 2 SOCIAL MEDIA ........ 52
TABLE 4.11: THREE FEATURE ANALYSIS SUPPORTING ACTOR 3 SOCIAL MEDIA ........ 54
TABLE 4.12: RESULTS CONCLUSION ........ 56
TABLE 4.13 ONE FEATURE ANALYSIS DIRECTOR AWARDS ........ 57
TABLE 4.14 ONE FEATURE ANALYSIS LEAD ACTOR AWARDS ........ 59
TABLE 4.15: ONE FEATURE ANALYSIS SUPPORTING ACTOR 2 AWARD ........ 61
TABLE 4.16: ONE FEATURE ANALYSIS SUPPORTING ACTOR 3 AWARD ........ 63
TABLE 4.17: TWO FEATURE ANALYSIS DIRECTOR AWARD ........ 66
TABLE 4.18: TWO FEATURE ANALYSIS LEAD ACTOR AWARD ........ 68
TABLE 4.19: TWO FEATURE ANALYSIS SUPPORTING ACTOR 2 AWARD ........ 71
TABLE 4.20: TWO FEATURE ANALYSIS SUPPORTING ACTOR 3 AWARD ........ 73
TABLE 4.21: THREE FEATURE ANALYSIS DIRECTOR AWARD ........ 76
TABLE 4.22: THREE FEATURE ANALYSIS LEAD ACTOR AWARD ........ 78
TABLE 4.23: THREE FEATURE ANALYSIS SUPPORTING ACTOR 2 AWARD ........ 79
TABLE 4.24: THREE FEATURE ANALYSIS SUPPORTING ACTOR 3 AWARD ........ 81
TABLE 4.25: RESULTS CONCLUSION ........ 83

# LIST OF ACRONYMS

CUST: Capital University of Science and Technology

MS: Master of Science

WOM: Word of Mouth

WOMM: Word of Mouth Marketing

SVM: Support Vector Machine

RF: Random Forest

MPAA: Motion Picture Association of America

IMDb: Internet Movie Database

ANN: Artificial Neural Network

DANN: Dynamic Artificial Neural Network

AMT: Amazon Mechanical Turk

# Chapter 1. Introduction

Every year, hundreds of movies are released by Hollywood with a belief that they can able to achieve valuable business from their investment[1] which depends on the success of their movie. However, few of them are able to get that much fame and appreciation from the audience and become a blockbuster. This achievement for the producers and directors is not an easy task, because they should keep in mind the taste and interest of the audience and viewers, everyone has their own taste according to the genre and the particular cast might not help. For example, different expectations occur in every individual related to some movie, few like comedy, however, others don't have interest in comedy and likewise some give priority to some specific actor or actress. So, this is highly complicated task for the directors to have a better idea about taste and interest of audience in order to achieve their goals. However, with the passage of time and advancement in the movie industry, we have successfully accessed very huge amount of data which could help us in better analysis of past trends and expected output in the future. So, these datasets can play key role for the future successful prediction for movies and can be a powerful addition to the existing system.

To get attention and interest of the audience and viewer, Hollywood industry needs to launch well entertaining movies which can inspire them. Now the question arises that what are the strategies which should be adopted to predict a movie to be successful among viewers. And can we be able to predict its success in the box office even before its production. These questions need to be answered so that we can get the maximum outcome from those movies. Jack Valente, quoted that "*No one can tell you how a movie is going to do in the marketplace. Not until the film open is darkened theatre and sparks fly up between the screen and audience*". Per this statement, this is not an easy job to predict a movie as successful before its release. Every director and producer wants to be successful on the box office and double their investments.

---

[1] https://goo.gl/AFetfw

In recent few years, researchers have concluded that people's attention and interest play an important role behind the movie's success. One can access this knowledge by using lots of online resources like IMDB (Internet Movie Database), user rating, comments and content available on the different social engine. We can also categorize the movies to most similar categories by utilizing past data of successful movies.

There are a lot of factors which plays an important role behind the success of a movie which includes their genre, casting actors and actresses, directors, producers, budget, marketing strategies, etc. So, we can able to analyze these factors at the pre-production stage to get strong outcomes and earnings from it. Movie makers are trying their best to get the attention of viewers by releasing outclass and entertaining movies. However, more work required to get the maximum outcome from these movies.

Number of existing productions have been considered which are used above strategies to predict success of movies. Some other factors are also considered to achieve strong outcomes such as tweets from people, interest of people for the trailers, their views and few other strategies for prediction of successful earning.

Another factor called Post-production also utilized to predict the success of a movie. It includes all the after shooting and after recording steps of production. If we considered existing work, a lot of research has been done on the post-production of the movie, which is an important segment for the movie success (Silva et al. 2014) and (Delen & Sharda, 2006). Also, WOM (word of mouth) has been used to visualize the overall accuracy. However, this method of post-production has some disadvantage, because prediction has no benefit after the investment. Therefore, the stakeholders are not able to recover their loss if the prediction goes wrong.

To accurately predict the success of movies, our research has analyzed different movie databases and by considering different contents from the social media. The important factor which is considered in our research is that we have utilized the pre-production as compared to post-production which is useless for the investors and stakeholders. Our research has explored previous 10 years' data of Hollywood (2005-2015) to accurately predict the successful outcome in box office for the year 2017. The feature sets of the movies have been shown in Table 3.2 in Chapter 3, which are used to classify the movie to be successful or flop. Different classifiers like Random forest, support vector machine, decision tree and Naïve Bayes have been considered in this research which we have trained using training data and have checked their result over the test data to see whether these trained models give an accurate prediction or not.

## 1.1 Purpose

This study's goal is to develop a forecasting model which makes prediction at an early stage of movie production which has practical value to investors or verdict maker in the film industry. All research papers generally give an idea related to one single algorithm proposed and its implementation. We hereby try to compare various algorithms on the ground of performance and efficiency.

## 1.2 Problem Statement

The vast majority of the literature evaluated the prediction of movie success after its release, however, at that time it's not valuable to the investor .Critical analysis of the literature surveys has led us to the following research gap:

1. Lead actor remains one of the popular parameter, however, their different awards like Oscar Awards, Golden Globe Awards, Venice Awards[2] and their social popularity of Facebook and Instagram has not been evaluated in the state-of-the-art research.
2. Similarly, director name was used in the past researches, however, Director's social popularity and awards won by him has not been analyzed.

**Research Question 1:** Which social media platform such as: Facebook, Twitter and Instagram can predict the success of a movie in a better way?

**Research Question 2:** Does the following awards: Oscar Awards, Golden Globe Awards and Venice Awards won by directors, lead and supporting actress/actor matter in the success of a Hollywood movie?

[2] http://www.worldfilmstreaming.com/

## 1.3 Scope

In this study, we proposed the evaluation of features to forecast box office success, promptly as a cast signs an agreement. This proposed forecasting time is the earliest prediction that was ever reported in the movie forecasting literature. The decision support system ranks cast by utilizing their performance of the last 10 years (2005-2015). In order to produce more accurate results, information based feature selection is also performed to select best subsets of features. After that; best set of feature will be proposed. This system tends to be dynamic tool, incorporating further data for real time adaption.

## 1.4 Significance of the Solution

It is clear that a movie's success is determined by different attributes. Prediction of a movie in pre-production stage will help studios to look into those attributes that movie a successful movie and also help stakeholders make decisions on whether to invest or not.

## 1.5 Dissertation Organization

The following sections will explain the structure and content of each chapter of this dissertation document.

Chapter 2: This chapter emphases on the techniques used for forecasting Box office revenue.

Chapter 3: defines the proposed methodology adopted in this research. It includes data collection techniques and other technique in order to conduct this research.

Chapter 4: present all the tables of experiment results and related discussions.

Chapter 5: this chapter concludes this dissertation and presents potential future research areas based on this research outcome. Further possibilities and approaches to further investigate this problem will also be discussed and recommendations will be presented at the end of the chapter.

## 1.6 Definitions and Abbreviations

### IMDb

The **Internet Movie Database** (abbreviated **IMDb**) is an online database of information related to films, television programs and video games, including cast, production crew, fictional characters, biographies, plot summaries, trivia and reviews.

### MPAA Rating

Motion picture Association of America (MPAA) is a body that assigns a rating to the movies. These ratings represent violence, sexual content, and language in a movie. There are 5 categories for each of the movies mainly R, PG, PG13, G and NR.

### Budget

Budget is the amount of resources that is used in the making of a movie. It is the total amount of money that is used in the whole making.

## 1.7 Conclusion

This chapter introduced this dissertation topic, the research problem, and the key objective and research questions to be answered by this research. The structure of this dissertation was also presented and the content of each chapter was explained briefly.

# Chapter 2. Literature Review

The main source of communication used now a day is social networking sites where people share their personal opinions, swap different perspectives and network at a very fast rate. Such social networking sites include Facebook (2 billion users), Instagram (700 million users), and Twitter (328 million users). Because of its convenience, speed and reach, online networking is quick at setting peoples' opinion, trends and agenda in entertainment, politics and technology.

It has always been an expensive and risky task to produce a successful movie. Determining models or algorithms have not been used by the production houses yet to foresee movie success. To minimize the loss, production houses have maximized the movie's budget. However, a successful movie cannot be guaranteed by the start power and maximized budget. It is the director's responsibility to produce a success movie when such huge amount of money has been spent. An assumed straight street to achievement has quite recently turned into a bended for creation studios. In any case, there is no such alternative with huge budget and star that ensures that a movie will be successful. This problem has given a good opportunity, especially to computer scientists. It all started with development of recommendation and predictive software to solve the problem.

Recommendation software became popular 1 million dollars prize was announced by Netflix for increasing the accuracy of their algorithm by ten percent[3]. However, development of predictive or forecasting models has not received that much attention. An interesting research domain in marketing and other disciplines has been provided by motion picture industry for scholars. The industry has high economic importance and is appealing to researchers because both rich data that cover the entire product lifecycle for many new products many unsolved "puzzles" have been provided by it (Jehoshua Eliashberg, Anita Elberse 2006; Elberse & Eliashberg 2003; Eliashberg et al. 2007; Eliashberg et al. 2000; Eliashberg 2000).

---

[3] http://goo.gl/AFetfw

For predicting the success of the movies, extensive study has been done by the WOM experts, neural network experts, econometricians as well as marketing professionals. In this paper, we reviewed forecasting models and a variety of analysis conducted by different researchers, however, our focus is on forecasting studies. In the case of a movie's success time prediction is of high importance, however, whole reviewed work falls in two major categories i.e. pre-production prediction and post-production prediction.  Our review also includes an overview of different domains and analysis of results of recent research studies.

Econometricians have been trying to find the contributing factors to predict the revenue of a movie which is normally achieved through linear regression analysis of movies data and examining the correlation between different determinants of revenue (Elberse & Eliashberg 2003). For analyzing the performance of sequels of different movies and to different to non-sequel movies, a significant number of studies has been carried out (Jehoshua Eliashberg, Anita Elberse 2006). A fascinating research area is analyzing that how does the success of a movie is affected by the stars. (Jehoshua Eliashberg, Anita Elberse 2006). The most accurate research project is MOVIEMOD (Eliashberg et al. 2000). This model predicts an error rate of ten percent, driven with the help of Markov Chains. This model has been used to evaluate the market before the movie distribution. For predicting the movie's success in different markets this model *(MOVIEMOD)* has played a vital role when it was tested.

Different social media signals, social media or word-of-mouth has been an important point for the researchers for evaluating the predictive value. To understand the consumer behaviors for different products, WOM has its own importance. Several studies have taken into account to understand if it is possible to predict the production performance with the data generated by consumer and how they are correlated with performance or sales. Results show that various social media signals play an important role in predicting the box office performance. A recent study by Shruti et al  has concluded that number of followers of an actor on twitter can be used as a predictive value for the movie success on box office however this predictive power is not present in Facebook (Shruti 2014).

However, very few researchers have tried to develop better models to predict the successes of movie on box office regardless of the unpredictable nature of movie's domain.

To employ statistical methods at post-production level, researchers have been trying to build forecasting models. The results were not accurate few years back (Delen & Sharda 2009; Sharda & Delen 2006) at pre-production level but a recent study has shown significantly better results (Moon et al. 2015). Most of the revenue comes from the first week of the movie's theatrical release as shown by experimental studies. This leads to another direction to predict movie's success in first week of release with very high accuracy. So, predicting movies success at a pre-production level becomes an interesting problem for the researcher.

## 2.1 Social Media

### 2.1.1 Facebook

Nowadays social media marketing is considered to be one of the top marketing strategies adopted by companies and individuals. According to Nielsen 2015, 84% trust is gained by users if their family, colleague or friends are using or referring a product and it is more likely that they will use or try that particular product. Similarly, 68% of people get influenced from other consumers making social media a great way to market a resource. Bulbul et al showed that 74% people follow a trend on social media (Bulbul & Shin 2014). Social media comments and likes work as recommendations for others and in a study it is said that 88% people trust such online reviews about a product. Among all social media networks Facebook is considered to be the top most which is joined by 20,000 users every second. It has 1.44 billion monthly users, 1.25 billion mobile users, 936 million daily active users. These facts and figures can tell that Facebook alone could be any companies marketing paradise which can be helpful to analyze and predict many things. Due to this influential property, we are using directors and first three main lead actors Facebook page likes to help us predict movies box office success or failure.

### 2.1.2 Twitter

One of the great promises of the social media revolution is that the ability to track people's interest in things in real time. Asur et al have used the chatter from Twitter to forecast box-office revenues for movies (Asur & Huberman 2010).The results are a fascinating insight into the power and limitations of Twitter. Popularity of directors, actors and movie on Twitter has great impact on predicting the financial success of a movie.

### 2.1.3 Instagram

Similarly, Instagram has fast become the preferred platform for sharing important and trivial moments alike with followers near and far. The social image-sharing platform now boasts 130 million active monthly users. Since its inception in 2010, Instagram has become a veritable channel for staying engaged with and informed by influencers. And for movie lovers, that means keeping up-to-date with studios about new releases and exclusive movie news.

## 2.2 Prestigious Award

An award is a prize which is given to a person for performing well in particular field. Oscar award (also known as Academy Award) is an award which is given by Academy of Motion Picture Arts and Sciences annually for the excellent performance. Very first Oscar award was presented in May 16, 1929. Golden Globe award is also a well-known award which is given by Hollywood Foreign Press Association for performing outstanding in films and television. The first Golden Globe award was presented in January 20, 1944. Venice Film Festival award was founded in 1932 in Venice, Italy. It is the oldest festival in the world.

## 2.3 Social Media Marketing Strategy

Different organizations have far and wide adapted social media, since then word of mouth marketing *(WOMM)* is contributing as an extensive source to predict movie's revenue on box-office. *WOMM* defined two major classes for movies i.e. data generated by critics and data generated by consumers. word of mouth marketing *(WOMM)* has been declared as best place to learn about consumer's choices and preferences as a result of various studies. Regardless of its significant predictive power, word of mouth marketing *(WOMM)* data could not be utilized by us as our research focuses at pre-production prediction. Since, word of mouth marketing *(WOMM)* is used as a tool to run campaigns, commercials and news sometimes which imply that collection of movies data is hard as nobody wants to advertise their product before launching. However, word of mouth marketing *(WOMM)* data have high prediction accuracy at post-prediction and post-release stages. Hence, the survey has been classified into two general categories such as: correlation based studies and regression analysis based studies. In following sections the critical review of both areas has been provided.

Generally, higher sales on box office are highly correlated to positive reviews but this may deny the general idea in many studies (Terry & Butler 2005). The impacts of negative reviews have stronger relationship with the low sales at box-office compared to positive reviews as suggested by the different experiments (Basuroy et al. 2003b; Basuroy et al. 2006; Basuroy et al. 2003a). However, revenue is not generally increased by positive reviews. This relationship has been observed in many other products sales especially in books sales. The relationship between performance of movies on box office and consumer generated data through WOM can be employed by correlation.

A research conducted by Krauss et al quite different but yet an impressive research (Krauss 2008). They focused on finding the relationship between online communities and financial success of movies. According to this study, the nominees for academy awards through the WOM data generated through communication of online communities were predicted. It also combined social network analysis with sentiment analysis and made very precise predictions. This research showed that community discussion on IMDB and probability of movie selection for the nominee for an academy award are positively correlated. They also found that if a movie is rigorously discussed then it could be successful but as not every movie get an intensive discussion but still succeeds. So, online discussions may not increase viewership in cinemas.

A recent study, using the volume of weblog about product to see the correlation with their sales, had tried to apply sentiment analysis to weblog data to see whether the results could be better than earlier or not (Mishne 2005). In the domains of movies the sentiment analysis of weblogs showed a quite higher correlation than volume only. Weblogs data have greater correlation with the financial performance of a movie when apply sentiment analysis. Experiments suggested that using the number of positive weblogs entries than using an only raw count will be better for evaluating the correlation between weblogs and financial performance at pre-release stage. Since, an accurate forecasting model cannot be built by the correlation on basis of these results. However, combining it with traditional factors can lead to an accurate predictive model for better forecasting.

Above study suggested a direction and someone took it seriously and tried to combine the traditional factors for predicting the movie success with sentiment analysis using data from news using Lydia (Lloyd et al. 2005). Lydia is a system for large-scale news analysis. The idea behind using the news sentiment analysis is that they carry extra weight than a normal statistical data for movies and it can be useful to predict the movie’s success on box office. Around 100 nationwide and local newspapers were given to the Lydia as input after this study (Zhang & Skiena n.d.). The following results were calculated, (i) grosses and articles are highly correlated than the reference, (ii) grosses and news references have more correlations with but have less correlation with the budget and (iii) positive references have higher correlation than the negative one with the grosses. It has proven that IMDB data and news analysis both have same effect for predicting a movie’s performance and especially for high grosses. However, using data form both the sources rather than only one source produced better results. Hence, it has been proved that combining the traditional features with sentiment analysis can have more predictive value than only one of them.

Furthermore, a study showed that blogs or news references of movies are highly correlated with their performance on box office (Zhang & Skiena n.d.). Another research study has been taken into account which has large number of features than usual i.e. 120. This study used all of the features related to the movies in blogs such as movies references, ranking and degree of reference (where was the movie name mentioned in the top paragraph of article etc.), references with respect to time (before release, after one, two, three or four weeks), sentiment analysis of positive references and combination of all above-mentioned features, which make it different from all other studies.

However, this study has distinctive research idea yet at the same time needed to agree with giants by using already used traditional features such as distributor name, genre and budget. These variables gross sale, critics' rating, and viewers' ratings were used for the correlation analysis. It could achieved $R^2$ 0.778 which is a quite better than the previous experiment which had only achieved 0.448. Even though combination of large numbers of features produced almost the same results i.e. blogs and news have same power to build a predicting model. According the results reported, however, it may be difficult to predict the movies sales after one of week of its theatrical release may be quite difficult. In this research, total number of movies was 197 which were not sufficient at all and that is why this was one of the major limitations of this study. Therefore, a big dataset is needed to testify the claim.

Influence of word of mouth marketing *(WOMM)* on different product is no more hidden now. However, its power varies from different domain which motivates the marketing managers to make better forecasting models using such medium. Now, the question is that which one is the better platform to achieve not only good sales but better sales given that they vary in power and require sensible selection. The following study Hyunmi et al has compared these different social media platforms like Twitter, Yahoo! Movies, YouTube and blogs to understand the above-mentioned question (Baek et al. 2014). How these four social media platforms affect the movies sales after release in 1$^{st}$, 2$^{nd}$, 3$^{rd}$ and 4$^{th}$ week? Twitter significantly has high impact of movies sales as compared to other three social media players, in first three weeks it showed the correlation around 0.632 but in forth week correlation was 0.492. So, calling a Twitter a mass media would not be wrong. The obvious reason behind this could be real time effects, wide spread trend of retweeting, combination of social network and expressing opinion in as few as possible words. On the other hand, unlike Twitter, Yahoo! Movies have shown low correlation 0.552 at early stages but higher 0.842 correlation at later stages of release. Above given results concluded that reviews are more significant at later stages. Comparison of blogs with Twitter showed that blogs have significant financial performance with the correlation of 0.634 at early stages but making a slight difference compared to Yahoo! Movies. Meanwhile, YouTube showed the same correlation in first three weeks 0.710 and 0.704 in the fourth week. Therefore, we can say that YouTube acts like both a social communication channel and a broad media.

In general, above-mentioned studies have tried to understand the correlation between different movies factors and revenue or sales but another class of researchers have been trying to understand the predictive power of different parameters. Therefore, a recent research has used the linear regression for predictive task. This study has no standard difference than the approach proposed by Zhang et al except by using the movie reviews before or on the release date of movies from seven different sources (Zhang & Skiena n.d.). These features have shown the correlation around 0.521 which did not make them significant factors of any movies success.

One of the most important parts of movies is Music. To make the movie more entertaining for its viewers, different soundtracks are produced. However, a couple of interesting research questions was not explored previously: can the number of searches for movies soundtrack establish any predictive power for the revenue in first week of theatrical release and subsequent weeks. Moreover, is there any difference between the existing and new songs track to influence the revenue (Lee & Jung 2014). Results have shown that soundtrack search volume of movie has an important correlation with revenue in the first week and it has same impacts in later weeks. Furthermore, the existing soundtrack highlights the relationship between the soundtrack search volume and revenue in initial stage but does not have the same effect in later stages.

Association of soundtrack search volume and movie revenues show that how movie revenue can be determined by different queries. Another idea is to include international audience through searches made at Google and combining them with traditional information like rating, number of screens and length of title of movie. Significant results have been shown that queries which includes the term related to the movies, affects the movies revenue (Lee & Jung 2014). However, length of a title, rating and number of screens has shown a slight difference. The approach had a limitation that no other learning algorithms than the linear regression were tried to get much better accuracy.

Exploring the effects of reviews on the revenue of movies is the only factor focused in above studies but none of them tried to investigate whether the reviews are influenced by the sales of movies or not. A study had demonstrated that user reviews are not significant when their indignity was taken into account (Duan & Keerthi 2005). Results concluded that higher rating did not require higher revenue on box office but frequency of posts extensively related to movies sales. This research is specifically based on the reviews of users on Yahoo! Movies and others sources of posting reviews have not been taken into account.

To evaluate the predictive power of different parameters, different research work has been done sometimes using the sentiment analysis, online rating, IMDB metadata and many others parameters to increase the accuracy of forecasting model. For example, Dellarocas et al found that to predict a movie revenue an opening weekend user ratings was highly significant (Dellarocas et al. n.d.). Accuracy can be increased by removing the imbalance reviews posted by males and females by equal weights. To post reviews for different products, Twitter has become as one of the best place for consumer and moviegoers also write reviews about movies at Twitter. Reviews posted by users on Twitter and their effect on revenue were studies by (Baek et al. 2014). This research concluded that movie revenue and volume of tweets are directly proportional to each other regardless of whether they are composed before two weeks or till the end of a film.

In another study by Asur et al tried to explore that how precise accuracy and better predictive forecasting model for the movies can be achieved through tweets (Asur & Huberman 2010). Vasu et al explored in his study that Twitter has more predictive power for movie success and a good accuracy can be achieved by basic sentiment analysis. However, this study had a limitation based on the dataset of only 1500 tweets per day, increasing the number of tweets may affect the accurate result (Jain 2013). All the previous discussion focused on those parameters and platforms which helped to generate WOM data. However, methodology is one of those important factors which did not get much attention. Question arises here that whether different mathematical models or learning algorithms make any difference in results.

Jean et al expended the above-mentioned approach by applying different algorithms to categorize the sentiment reviews on Rotten Tomatoes, annotated by AMT (Amazon Mechanical Turk) (Wu n.d.). In their research they concluded that Multinomial Naïve Bayes (MND) and Support Vector Machine (SVM) have much better accuracy than neural network. Hence, from all the previous discussions it is concluded that to learn the consumer behavior or developing accurate forecasting models, word of mouth marketing *(WOMM)* has always been an important source. However, an important fact about word of mouth cannot be ignored that significant accuracy cannot be achieved by WOM before the release or pre-production of movies as word of mouth data becomes available only after the release of a movie. Therefore, our main focus will be on different forecasting models (pre-production or post-production) and their different categories.

## 2.4 Forecasting of Motion Picture Revenue

Forecasting models are categorized into further two classes to predict the success of a movie based on the timing of inputs and this classification is divided according the methodologies used. To forecast the revenue of a movie Regression, Bayesian and Artificial Neural Networks models have repeatedly been used by researchers. Accurate prediction of a movie success can be calculated using different forecasting studies using more parameters, number of receipts, critics and user reviews.

Neelamegham et al investigated remarkable study using Bayesian methods to predict the revenue (Neelamegham & Chintagunta 1999). Their models successfully showed significant improvement of accuracy of Sawhney et al from 45% to 71% at pre-release stage (Eliashberg et al. 2000). Artificial Neural Networks have been extensively used in the domain of forecasting over the years. Neural Network has established superiority with the counterparts such as regression based models, discriminate analysis and support vector machine (Delen & Sharda 2009; Sharda & Delen 2006; Zhang et al. 2009). An interesting study about the model used for the post-prediction forecasting has been discussed in the section below which will show that post-production forecasting studies have better results than the pre-production forecasting studies.

### 2.4.1 Post-Production Forecasting Models

Sharda et al conducted the revolutionary research in post-production prediction (Delen & Sharda 2009; Sharda & Delen 2006). By defining ten different revenue classes, they converted the forecasting problem into classifications. The basic idea behind this research was to evaluate the neural network expertise to predict movie's revenue. Pre-release prediction was the contributing factor of this research. Results demonstrate that by using neural network 37% of accuracy can be predicted. Although the overall accuracy was not up to that point but when compared with others, neural network has better results than discriminate analysis, logistic regression and classification tress. Extension can be done to this model using different parameters and features set and better training and testing can be done to increase accuracy of model which attracts researches towards this domain.

Artificial Neural Network has been significantly defined in the above-mentioned study, however, by changing the algorithm may help to increase the accuracy. Zhaing at el (Zhaing, Luo and Yang, 2009) used Black Propagation Neural Network in their study which employed Multi-Layer Perception to predict movies performance on box office (Zhang et al. 2009). Six different classes of movies were defined according to their revenues. Significant predictive model was achieved when the number of layers for training the model went through several experiments. Data can be split into training and testing data using 10-Fold cross validation. It has become a standard methodology. However, changing the split ratio may help sometimes and same happened during this study. Optimal performance was achieved by splitting the data in 6-Fold. Significant improvement of 68.1% from 37% for predicting the performance of movie has been shown by Back Propagation neural network. Hence, Back Propagation neural network won the race in building better predictive model for movies when compared with Multilayer perceptron neural network.

Most of the classifying algorithms have one common assumption that every movie is independent of all other movies. However, this assumption is wrong for revenue prediction problem of movies. Parimi et al identified a graph structure. There are several reasons for which movies can be linked (Parimi & Caragea 2013). For example, if movies have same lead actor/actress, director, genre, and sequel or may have the same releasing date. Reputation of same leading role or a director may help to make a movie successful. For example, a movie will have better performance if its director has a good repute as compared to that movie which does not have a reputable director. Thus, to improve the accuracy, the idea of dependency between different movies has been proved correct. Therefore, better predictive models can be built using dependency between different features.

As we discussed earlier that changing the learning algorithm may help to increase the accuracy, Parimi et al  investigated the same thing when they considered the dependency between the movies to improve the accuracy of an algorithm (Parimi & Caragea 2013). Classification has many algorithms to achieve the results as discussed above. Performance of Decision Trees and Artificial Neural Networks is almost same. This study concluded that decision learning is an appropriate algorithm for building the predictive model for Hollywood gross income.

We have seen that how word of mouth marketing can influence movies sales. Recently, a group of researchers developed a web-based decision support system for managers to predict the future of products. The application used a non-traditional forecasting model by taking more than one independent experts to make one variable decision support system. More than one prediction models are used to predict the financial performance of movies namely ANN, Decision Tree, Logistic Regression and Discriminate Analysis. 849 movies data was collected from IMDB on which models were trained. Large number of users accepted this system. Since most of the users were college students which was immature decision. They may try to test their system for the real decision makers.

Kulkarni et al  in their study they explored that how data about the searches of customer related to a particular product will help to understand customer's preferences (Delen & Sharda 2009). For better advertising and sale campaigns this data is quite precious for managers. This idea has given a direction to build a forecasting model for pre-release search pattern of consumer related to particular product at pre-release stage and in this case especially for the movies. These results concluded that higher accuracy can be achieved by word of mouth data. However, word of mouth data becomes less practical option when we want to predict the movies success at pre-release or post-production stages.

We have studied that how the concept of search volume related to a particular product can help us to make prediction and appropriate for online review and rating. These concepts provide an opportunity for organizations to make better strategies/plans for future. In the study of  Moon et, they investigated that how different factors like box-office data, star rating, search volume of movies titles and many other factors have imperative effect on sale of movies tickets (Moon et al. 2010). Results of this research showed that by using box office and other external (mentioned above) 9% of accuracy would have been improved. Performance of ANN had a unique but in later weeks SVM became a bit unstable. The movies data of this study was related to Korean Movies therefore, applicability is limited.

Use of big data generated online by consumer has become a primary interest in computation business intelligence to understand collective opinions. Mestyan et al identified that naïve application would be better to predictive power and consumer reaction to new different products (Mestyán et al. 2013). The same concept is applicable for predicting the movies revenue by using data about editors and viewers of a movie page on Wikipedia. For this, a number of views, the number of users, edits and theaters were multivariate linear regression analysis used as explanatory variables and was applied to calculate the coefficient of determination. The results demonstrated that a number of theaters in which film was released turned out the best factor to predict the success. Accuracy about 0.98 coefficient of determination was achieved which was far better than the Twitter which was 0.94. However, it would not be wrong to say that Twitter based model had showed prediction of high accuracy just before the release. This model is quite simple than other models that revealed the effect that most of the editors on Wikipedia are determined followers of movie industry. They not only write things but they also review information about move before its release on theaters which make the Wikipedia a more powerful in predicting success.

Above research clearly demonstrated that how Wikipedia can help to build high accuracy predictive model as compared to many other approaches. The idea of checking the application of model in other countries is quite interesting. De Silva et al in a recent research assessed their model for 325 movies released in the United States against movies in five other countries (Silva et al. 2014). Results showed that previously reported models which involve Wikipedia have better accuracy rate of prediction. On other hand, number of views on movies Wikipedia page are not strong enough to predict the success in opening week in other countries like Japan, United Kingdom and Germany. People of US and Australia can view the Wikipedia when they search for a best movie option before watching, but this may not happened with people of those countries. They use some other sources for this sake. There is another unsolved question that how this model was more appropriate for Australia as compared to other countries. Therefore, for building the accurate predictive model for the movie success in Japan, UK and Germany, someone should consider other parameters along with the Wikipedia page views.

Building a prediction model is indeed very challenging task but also an interesting problem. This is why different researchers from different domains tried to build better an accurate predictive model. However, all these models have certain limitations. In the study of Sharda et al , they had tried information infusion approach and combined more than one learning algorithms and dataset with more features (Delen & Sharda 2009; Sharda & Delen 2006). However, result were not quite promising and had 56% of combined accuracy. The accuracy can be improved by adding all those features which were not used in this model. Hence, possibility of improving the model becomes an open opportunity for the future experiments.

M. Assady et al tried to combine the power of both machine learning and visualization to obtain an accurate predating user rating and revenue (Assady et al. 2013). Two different models long-term rating and short-term rating with the accuracy of 0.45 and 060 respectively have been developed. Long-term rating produced low error rate and made it bias. For future improvement in long-term rating, different methods may be dealt in future.

Visual Analytics Science and Technology Challenge 2013 encouraged many researchers to make better predictive model with the help of visual analytics. According to Philipp et al , only limited number of studies has been reported in this domain(Wang et al. 2001). Prediction of movies rating and revenue has been achieved by using neural network which is an appropriate selection of algorithm. Results demonstrate that only for predicting viewer rating forecasting was quite accurate. Movies success can be predicted by making enhancements in proposed models/methods. Visual analytics did not get much attention in the domain of movies success prediction. The VAST challenge 2013 has invited many researchers to reveal facts that human brain can extract information and analyze complex data quickly. Huge amount of data related to product are being collected. This data helps to understand the complicated patterns, association and prediction.

### 2.4.2 Pre-Production Forecasting Models

Majority of the research studies focused on building a forecasting model at post-production or post-release stage and those models were able to make precise prediction with significant accuracy. However, data for these highly accurate models were collected from word of mouth data which are available after the release of a movie or the first weekend of it. Unfortunately, prediction at that time is of no use as different stockholders have invested their money on it. So, prediction after the release is not significant to avoid the loss. Therefore, the following study which focused on pre-production predictions of movie’s revenue on box office was published (Ghiassi et al. 2015). More than 80% of predictions were made using proposed model which becomes the benchmark for other studies. By using these results, importance of the variables and learning algorithm were highlight that helped in building the predictive model. Accuracy of both the testing and training data were compared.

Overview of significant researches conducted in past has been presented above. Different initial work and some highly contributing research ideas have been studied. Plenty of work had been done in predicting the movie's revenue and still number of different competitions had been arranged to motivate researchers to do their research studies in this domain[4] and huge amount of money was offered to winners (Assady et al. 2013).

Several methodologies were part of our research discussed above; each methodology has its own importance and deals with specific nature of data related to movies. For example, different research studies have been explored that use word of mouth data to test correlation of different attributes with movie revenue in first weekend of movie (Basuroy et al. 2003b; Asur & Huberman 2010). To check that how explanatory performed on their proposed models, correlation analysis has been used and evaluated using determination coefficient. On the other hand, other researchers are still using word of mouth data to build accurate predictive models using different parameters with regression analysis and evaluate how Means Squared Error *(MSE)* can be used (Dellarocas et al. n.d.; Zhang et al. 2015).

This is not an end of the story, quantitative studies have been explored to predict the revenue of movies using different traditional parameters such as start value, competition, seasonality, etc. but those models tend to work for forecasting at pre-release or post-production level that successful research study which used different dimension to predict the revenue at pre-production level (Ghiassi et al. 2015; Delen & Sharda 2009; Sharda & Delen 2006). Our research also based on pre-production prediction for Hollywood movies considering all previous research in this particular domain.

[4] http://goo.gl/AFetfw

# Chapter 3. Methodology

In the previous chapter, critical review has been represented which showed that the majority of studies have tried prediction of movie success class at post-production level. However, none of them have used features which are available from previous success or failure of actors, actresses, or directors for example lead actors and directors' different awards like Oscar Awards, Golden Globe Awards, and Venice Awards. Similarly their social popularity of Facebook and Instagram has also not been evaluated in the previous researches.

Based on the research gap identified in the previous chapter, this thesis formulates the following research questions to be evaluated:

**Research Question 1:** Which social media platform such as: Facebook, Twitter and Instagram can predict the success of a movie in a better way?

**Research Question 2:** Does the following awards: Oscar Awards, Golden Globe Awards and Venice Awards won by directors, lead and supporting actress/actor matter in the success of a Hollywood movie?

In order to evaluate these research questions, we have adopted an appropriate methodology as shown in the Figure 3.1. The first step was to collect the comprehensive dataset of movies. Data was collected from four different sources like IMDb, Facebook, Twitter and Instagram and has been discussed in details in the Section 3.1. The detail of the features used in this research has been discussed in details in the section 3.2. After collecting the raw data, pre-processing step has been discussed in the Section 3.3.

After pre-processing, number of models has been trained and tested using particular experimental settings as detailed explanation has been provided in the Section 3.4.

## 3.1 Data Collection

In this thesis, Hollywood movies data were used. We have chosen Hollywood field because they produce a variety of movies. Data was collected manually from four well-known sources IMDB, Facebook, Instagram and Twitter. Firstly, movies' title, director names, actor 1, 2 and 3 names, budget, IMDB rating and different awards won by them such as Oscar awards, Golden Globe awards and Venice awards were collected from IMDB. The followers of directors and actors were collected from Instagram and Twitter respectively. Nine ranges of revenue are considered to construct the dataset which become nine classes to be considered by the classifier as shown in Table 3.1.

Table 3.1 Output Class Revenue Ranges

| **Class** | **Revenue Range (in $ millions)** |
|---|---|
| A *(blockbuster)* | 200 + |
| B | 150 to 200 |
| C | 100 to 150 |
| D | 65 to 100 |
| E | 40 to 65 |
| F | 20 to 40 |
| G | 10 to 20 |
| H | 1 to 10 |
| I *(flop)* | < 1 |

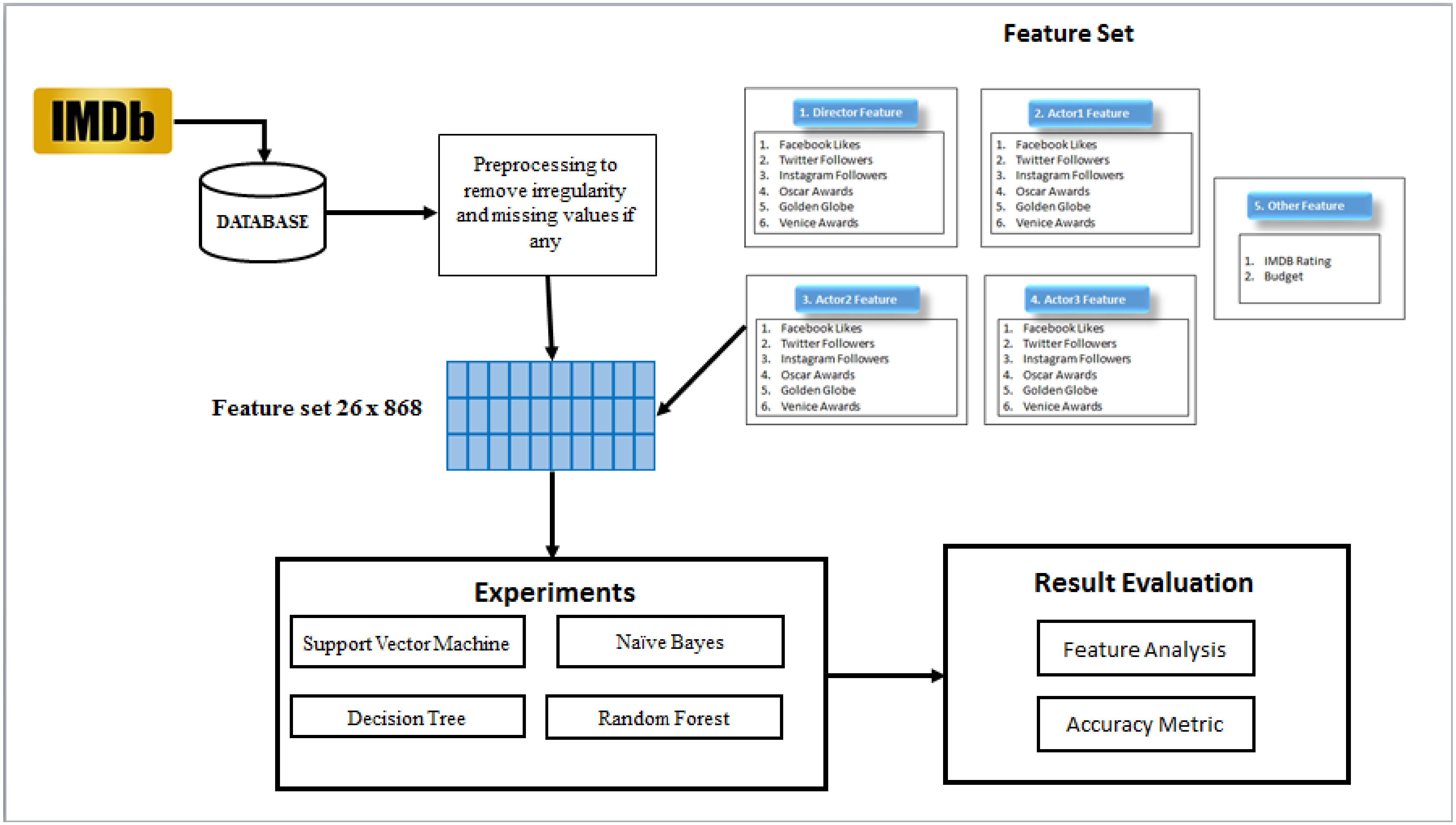


Figure 3.1 Methodology and Feature Set

## 3.2 Feature Set

For this research, 24 unique features which contain last 10 years data were used. Table 3.2 shows the feature set calculation. Features such as Facebook, Twitter and Instagram have been used because in previous researches none of the researchers have used them for directors and actors in their research. Similarly awards are given to praise the performance of individual in the particular field. Awards related to movies are given on the basis of movies critics and audience opinions. If both the critics and audience have same and strong opinion, then the award is given to that person. It means that this type of features may help in predicting the success of a movie. In this thesis, we have used all these features to forecast the success of a movie. Other than these features, budget and IMDB rating of a movie is also obtained from IMDB.

Table 3.2 Feature Set

| S No. | Features |
|---|---|
| 1 | Director_Twitter_Followers |
| 2 | Actor1_Twitter_Followers |
| 3 | Actor2_Twitter_Followers |
| 4 | Acto3_Twitter_Followers |
| 5 | Director_Instagram_Followers |
| 6 | Actor1_Instagram_Followers |
| 7 | Actor2_Instagram_Followers |
| 8 | Actor3_Instagram_Followers |
| 9 | Director_Facebook_Likes |
| 10 | Actor1_Facebook_Likes |
| 11 | Actor2_Facebook_Likes |
| 12 | Actor3_Facebook_Likes |
| 13 | Director Oscar Award |
| 14 | Director Golden Globe Award |
| 15 | Director Venice Award |
| 16 | Actor1 Oscar Award |
| 17 | Actor1 Golden Globe Award |
| 18 | Actor1 Venice Award |
| 19 | Actor2 Oscar Award |
| 20 | Actor2 Golden Globe Award |
| 21 | Actor2 Venice Award |
| 22 | Actor3 Oscar Award |
| 23 | Actor3 Golden Globe Award |
| 24 | Actor3 Venice Award |

## 3.3 Pre-Processing

After collecting the data, pre-processing has been applied on the data to check whether it contained null values or not. Our data contained lots of missing values as shown in Figure 3.2.

The highlighted part showed that each row represents the record of one movie. These rows contain some missing values of directors and actors. The reasons behind these missing values are:

- Verified accounts of some directors, lead actors and supporting actors on Facebook, Twitter and Instagram are not available. Therefore, it was difficult task to obtain the number of followers.
- Hollywood industry also produces animated movies which do not have supporting actors.
- Moreover, it is quite possible that directors or actors may not win any award.

This may result in incorrect evaluation. Therefore, those rows which contained null values have been deleted to achieve comprehensive dataset and to avoid incorrect results. After that, comprehensive dataset with 24 features was obtained.

| Director Name | Director Facebook Likes | Director Twitter Likes | Director Instagram Likes | Actor1 Name | Actor2 Name | Actor3 Name |
|---|---|---|---|---|---|---|
| Chris Kentis | 9 | 0 | 16059 | Eric Sheffer Stevens | Julia Taylor Ross | Adam Trese |
| Alex Kendrick | 589 | 5067 | 0 | Ben Davies | Alex Kendrick | T.C. Stallings |
| Colin Minihan | 6 | 97027 | 12 | Mackenzie Gray | Sean Rogerson | Ashleigh Gryzko |
| Tanner Beard | 531 | | 107 | William McNamara | Kevin Alejandro | Glenn Morshower |
| Vera Farmiga | 0 | 0 | 212083 | Donna Murphy | Bill Irwin | Michael Chernus |
| Morgan Spurlock | 293 | 165128 | 17112 | Quentin Tarantino | J.J. Abrams | Donald Trump |
| Deryck Broom | 11 | 140 | 0 | Vic Mignogna | Omar Benson Miller | Anupam Kher |
| Léa Pool | 4 | 848 | 142287 | | | |
| Laurent Bouhnik | 0 | 357265 | 6993439 | Déborah Révy | Johnny Amaro | Yassine Azzouz |
| Justin Thomas Ostense | 0 | 0 | 263794 | Michael Berryman | Kristin Booth | Michael Eisner |
| Gareth Evans | 338 | 347 | 8254 | Iko Uwais | Yayan Ruhian | Donny Alamsyah |
| Andrew Erwin | 10 | 770 | 0 | Rachel Hendrix | Robert Amaya | Jason Burkey |
| Sean Durkin | 46 | 5576 | 0 | Christopher Abbott | Julia Garner | Brady Corbet |
| Stefan C. Schaefer | 0 | 380 | 0 | Nicole Beharie | Marlene Forte | Reg E. Cathey |
| Kat Coiro | 28 | 6252 | 370153 | Justin Kirk | Geoff Stults | Kristen Johnston |
| Maryam Keshavarz | 32 | 114 | 0 | Sarah Kazemy | Reza Sixo Safai | Sina Amedson |
| Mariette Monpierre | 0 | 347 | 0 | Stana Roumillac | | |
| Asghar Farhadi | 0 | 0 | 4328 | Shahab Hosseini | Leila Hatami | Peyman Moaadi |
| Leslie Small | 15 | 988 | 11 | Isaac C. Singleton Jr. | Larry King | Jeanette Branch |
| Ben Wheatley | 214 | 32693 | 0 | MyAnna Buring | Ben Crompton | Neil Maskell |
| Ti West | 243 | 29879 | 7603 | Lena Dunham | Jake Ryan | Pat Healy |
| U. Roberto Romano | 6 | 7068 | 79607 | | | |
| Matt Walsh | 490 | | 28588 | Joe Lo Truglio | Matt Jones | Abby Elliott |
| Joseph Dorman | 0 | 8876 | 2517415 | Rachel Dratch | Peter Riegert | Jason Kravits |
| Jack Heller | 0 | 0 | 0 | Shaun Sipos | Katherine Waterston | Christopher Denham |
| Jack Perez | 19 | 216402 | 2048538 | Kevin Corrigan | Barry Bostwick | Ahmed Best |
| Drake Doremus | 52 | 7432 | 2001 | Jennifer Lawrence | Charlie Bewley | Finola Hughes |

Figure 3.2 Missing values

## 3.4 Experiments and Evaluation

This thesis comprehensively evaluates the following features:

1) Features related to social popularity and awards won by directors
2) Features related to social popularity and awards won by Lead Actor

3) Features related to social popularity and awards won by Actor 2
4) Features related to social popularity and awards won by Actor 3

To evaluate these features and to identify which feature can produce good results, we need to look into literature to identify which classifiers produce better results. After critical analysis we have evaluated that these four classifiers have produce better results than others. These classifiers are Random Forest, Naïve Bayes, Support Vector Machine and Decision Tree. First of all, two categories of classes have been defined. In first category, nine revenue classes (class “A” to class “I”) are used and in second category only two classes Successful class (mapped as class “A” to class “D”) and Un-successful class (mapped as class ”E” to class “I”) are used.

We have used split ratio approach which is one of those approaches which mostly practice in machine learning. In this model, a classifier will be trained on dataset from 2005 to 2014, and will be tested on 2015. Then the evaluating parameters Precision, Recall and F-Measure were calculated to evaluate the accuracy of each classifier as shown in equation (1), (2) and (3).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (1)$$

Precision is calculated for those forms or forms which are correctly selected.

$$Recall = \frac{True\ Positive}{True\ Positive + True\ Negative} \qquad (2)$$

Recall is calculated for those form or forms which are successfully selected. Furthermore, the f measure is calculated for these respective classes and simplifies the results.

$$F\ measure = 2 * \frac{Precision\ .\ Recall}{Precision + Recall} \qquad (3)$$

# Chapter 4. Results

This chapter presents the detailed results of proposed methodology discussed in the previous chapter. The obtained results and their significance and the influence of social media and different awards have been discussed. Dataset collection statistics has been discussed in the first section. Results have been divided into two categories; the first category discusses the F-measure of single feature in the dataset using output classes (classes ranging from class “A” to class “I” as explained in Table 3.1) and the second category discuss the F-measure with only 2 classes A and B. in this scenario, classes “A” to class “D” have been mapped on Successful class and from class “E” to class “I” have been mapped to Un-successful class. In other words, successful movies are those who have generated above average revenue, and Un-successful class represents those movies which have generated. At the end, social media and awards’ impact on the success of a movie have been discussed.

## 4.1 Data Statistics

We have manually collected the data of Hollywood movies for last 10 years (2005-2015). In these years number of movies with different genre such as Comedy, Horror, Action, Romance, Animated etc. has been produced. Collecting movies of such different genres makes data valuable. Many important movies information were considered from IMDb website. For example, movie title, year, director name, lead actors and supporting actors’ names etc. Other information like number of followers and Facebook page likes were collected from Twitter, Instagram and Facebook. This process took almost 3 hours to finish. In the end, total numbers of 2000 movies were obtained. There are 2000 Directors and Lead Actors (Actor 1), 1992 Actor 2 and 1988 Actor 3.

## 4.2 Pre –Processing

Initially, total number of movies was 2000 which was collected from different sources such IMDb, Facebook, Twitter and Instagram. In each row, record of one movie is shown. However, this data contained some missing values as well. Those rows which have missing values for example, some movies are animated and they do not have supporting actors.

Similarly, directors or actors verified accounts have not been created therefore; their followers have not been collected. For these reasons, records of all such movies were deleted, so that our results do not become biased and comprehensive dataset has been obtained. Almost 1100 of data was deleted which contained null values. After performing cleaning, comprehensive dataset of 868 movies has been collected.

## 4.3 Evaluation

This section presents the detailed evaluation of all features discussed in the chapter 3. This thesis has raised two research questions, the answer of first research question has been discussed in the section 4.4 and the answer of the second research question has been discussed in the section 4.5.

## 4.4 Social Media Impact on Movie Success

First research question was:

**Which social media platform such as: Facebook, Twitter and Instagram can predict the success of a movie in a better way?**

To evaluate this question, we have comprehensively evaluated three social platforms such as number of followers of actors and directors at: (1) Facebook, (2) Twitter, and (3) Instagram. We have evaluated each of these as independently as well as by combining in possible ways. For example, in first type of evaluation, each feature was evaluated, in second type of evaluation, two features were combined to compare the results, and in third evaluation, all three features have been collectively checked. F-measure of each feature of director and actors has been calculated using four different classifiers. F-measure of combinations of Facebook, Twitter and Instagram has been calculated to check which feature performs best. Furthermore, there are two types of evaluations in each case. In first case, F-measure of each feature has been calculated for classes from class “A” to class “I” then the F-measure of two classes Successful and Un-successful has been calculated.

### 4.4.1 One Feature Analysis

**(A) One Feature Analysis for Director**

The results of one feature analysis for director have been shown in this section.

Table 4.1 One Feature Analysis Director Social Media

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook | 0.715 | 0.74 | 0.724 | 0.2 | 0.237 | 0.206 |
| | Twitter | 0.697 | 0.705 | 0.701 | 0.192 | 0.197 | 0.19 |
| | Instagram | 0.724 | 0.751 | 0.733 | 0.263 | 0.277 | 0.256 |
| **2. Naïve Bayes** | Facebook | 0.738 | 0.775 | 0.7 | 0.17 | 0.347 | 0.193 |
| | Twitter | 0.591 | 0.769 | 0.668 | 0.148 | 0.324 | 0.193 |
| | Instagram | 0.688 | 0.295 | 0.231 | 0.163 | 0.092 | 0.066 |
| **3. Support Vector Machine** | Facebook | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Twitter | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.169 |
| | Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Facebook | 0.591 | 0.769 | 0.668 | 0.21 | 0.335 | 0.239 |
| | Twitter | 0.591 | 0.769 | 0.668 | 0.226 | 0.254 | 0.215 |
| | Instagram | 0.591 | 0.769 | 0.668 | 0.287 | 0.289 | 0.242 |

With all classes' *Random forest* classifier *Instagram* achieved highest F-measure of 0.256. *Facebook* achieved F-measure of 0.206 and *Twitter* achieved the lowest F-measure of 0.19. *Facebook* and *Twitter* achieved same F-measure with *Naïve Bayes* i.e. 0.193. However, *Instagram* achieved lowest F-measure of 0.066. *Twitter* achieved F-measure of 0.169 while *Facebook* and *Instagram* have achieved the F-measure of 0.168 with *Support Vector Machine*. With *Decision Tree, Instagram* obtained highest F-measure of 0.242. *Facebook* achieved 0.239 F-measure and *Twitter* achieved the lowest F-measure of 0.215 as shown in the Figure 4.1.

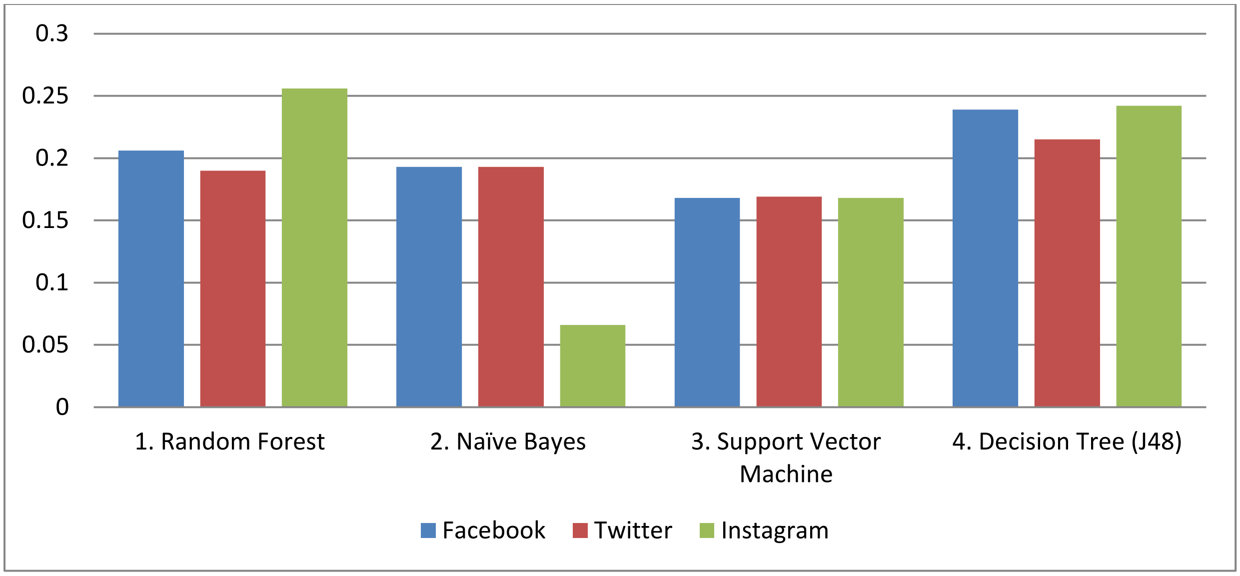


Figure 4.1 Director One Feature Analysis Class A to Class I

**Findings:** When single feature is used, the feature *Instagram* obtained the best result by scoring F-measure of 0.256 with *Random Forest Classifier*.

Results of two classes are shown in the Figure 4.2. With *Random forest* classifier *Instagram* achieved the highest F-measure of 0.733 while *Facebook* achieved F-measure of 0.724 and *Twitter* achieved the lowest F-measure of 0.701. *Twitter* achieved the highest F-measure of 0.688 with *Naïve Bayes*. However, *Instagram* has achieved F-measure of 0.231 and *Facebook* achieved the lowest F-measure of 0.7. *Facebook*, *Twitter* and *Instagram* obtained the same result by scoring F-measure of 0.668. With *Decision Tree*, all these social media achieved the same F-measure of 0.668.

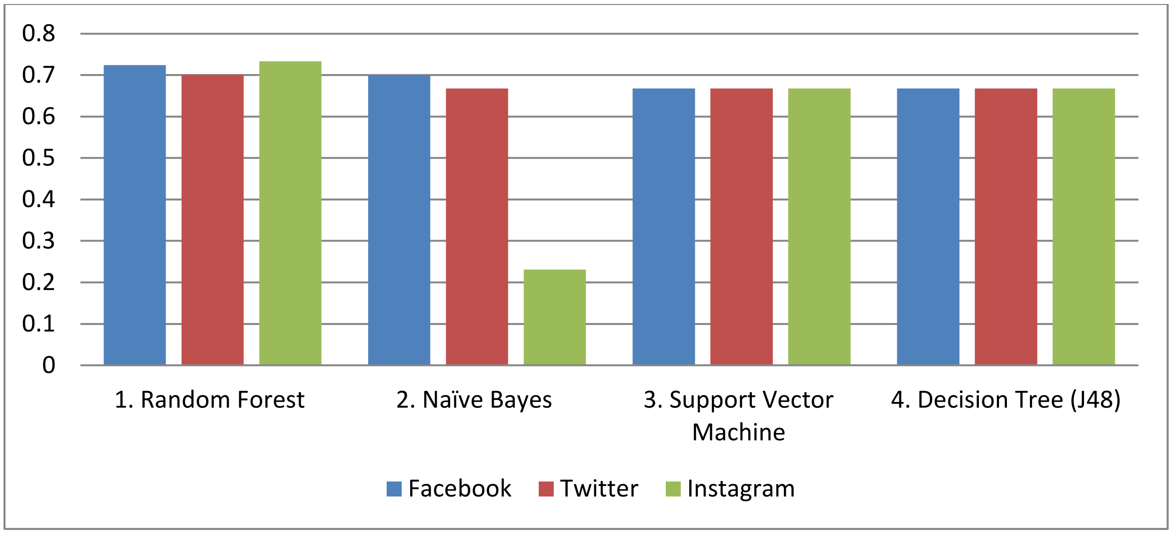


Figure 4.2 Director One Feature Analysis Successful and Un-successful Classes

**Findings:** The feature *Instagram* obtained the best result by scoring F-measure of 0.733 with *Random Forest Classifier*.

### (B) One Feature Analysis for Lead Actor 1

The results of one feature analysis for lead actor 1 have been shown in this section.

Table 4.2: One Feature Analysis Lead Actor Social Media

| **Classifier** | | **Two Output Classes** | | | **All Classes** | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook | 0.667 | 0.723 | 0.686 | 0.2 | 0.237 | 0.206 |
| | Twitter | 0.711 | 0.751 | 0.721 | 0.192 | 0.197 | 0.19 |
| | Instagram | 0.715 | 0.74 | 0.724 | 0.263 | 0.277 | 0.256 |
| **2. Naïve Bayes** | Facebook | 0.591 | 0.769 | 0.668 | 0.17 | 0.347 | 0.193 |
| | Twitter | 0.709 | 0.769 | 0.679 | 0.148 | 0.324 | 0.193 |
| | Instagram | 0.59 | 0.763 | 0.665 | 0.163 | 0.092 | 0.066 |
| **3. Support Vector Machine** | Facebook | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Twitter | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.169 |
| | Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Facebook | 0.591 | 0.769 | 0.668 | 0.21 | 0.335 | 0.239 |
| | Twitter | 0.591 | 0.769 | 0.668 | 0.226 | 0.254 | 0.215 |
| | Instagram | 0.591 | 0.769 | 0.668 | 0.287 | 0.289 | 0.242 |

With *Random forest* classifier *Instagram* achieved highest F-measure of 0.256. *Facebook* achieved F-measure of 0.206 and *Twitter* achieved the lowest F-measure of 0.19. *Facebook* and *Twitter* achieved same F-measure with *Naïve Bayes* i.e. 0.193. However, *Instagram* achieved lowest F-measure of 0.066. *Twitter* achieved F-measure of 0.169 while *Facebook* and *Instagram* have achieved the F-measure of 0.168 with *Support Vector Machine*. With *Decision Tree, Instagram* obtained highest F-measure of 0.242. *Facebook* achieved 0.239 F-measure and *Twitter* achieved the lowest F-measure of 0.215 as shown in Figure 4.3.

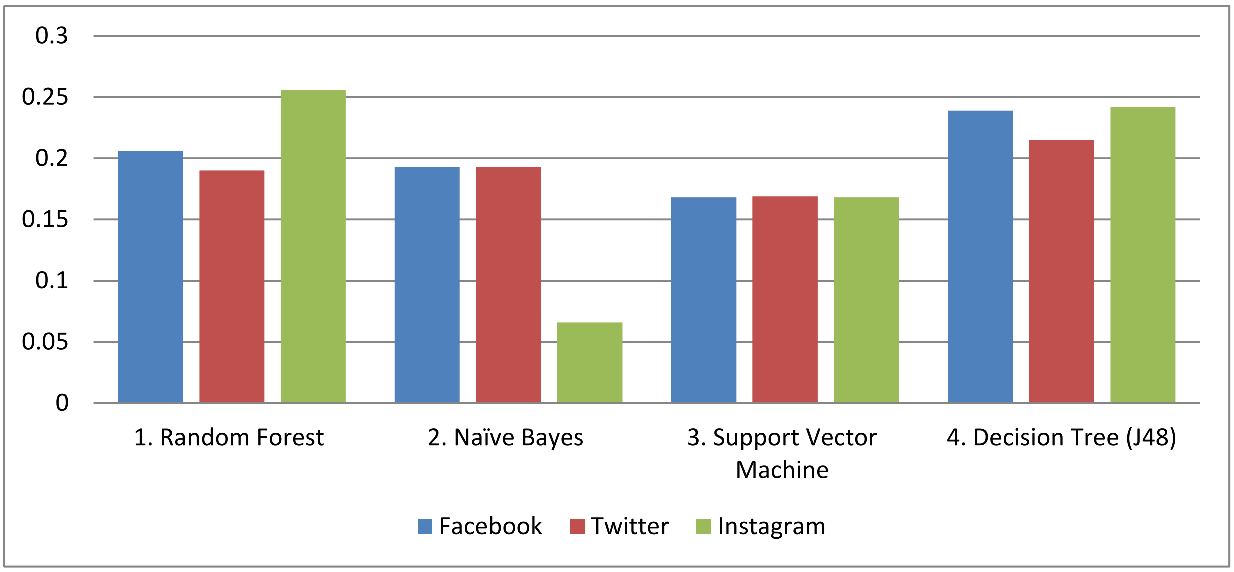


Figure 4.3 Actor one- One Feature Analysis Class A to Class I

**Findings:** With all classes, the feature *Instagram* obtained the best result by scoring F-measure of 0.256 with *Random Forest Classifier*.

Figure 4.4 show the results of two classes. With *Random forest* classifier *Instagram* achieved the highest F-measure of 0.724 while *Twitter* achieved F-measure of 0.721 and *Facebook* achieved the lowest F-measure of 0.686. *Twitter* achieved the highest F-measure of 0.679 with *Naïve Bayes*. However, *Facebook* has achieved F-measure of 0.668 and *Instagram* achieved the lowest F-measure of 0.665. *Facebook*, *Twitter* and *Instagram* obtained the same result by scoring F-measure of 0.668. With *Decision Tree,* all these social media achieved the same F-measure of 0.668.

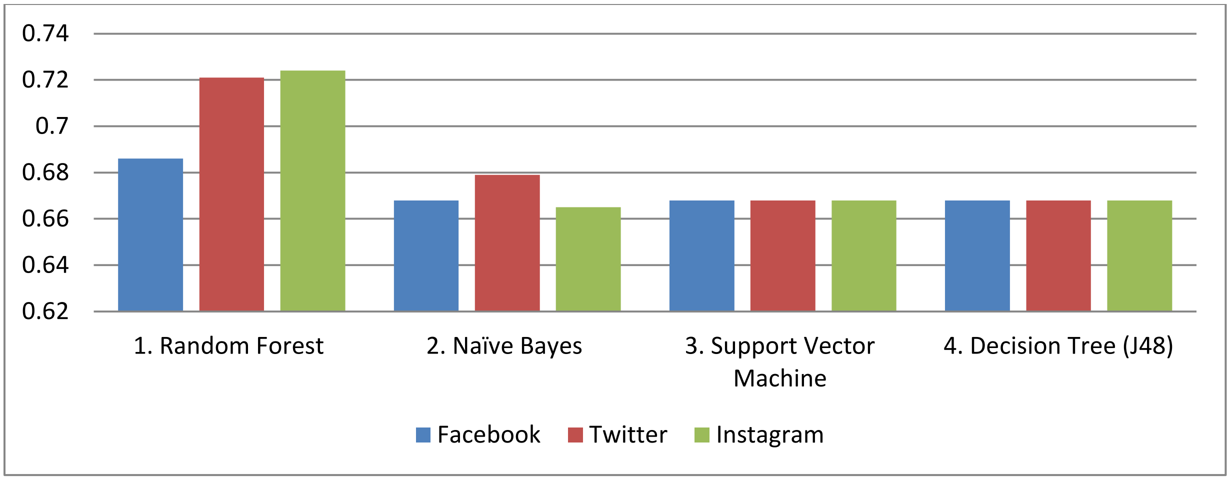


Figure 4.4 Actor one- One Feature Analysis Successful and Un-successful Classes

**Findings:** *Instagram* obtained the best result by scoring F-measure of 0.724 with *Random Forest Classifier*.

**(C) One Feature Analysis for Supporting Actor 2**

The results of one feature analysis for supporting actor 2 have been shown in this section.

**Table 4.3: One Feature Analysis Supporting Actor 2 Social Media**

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook | 0.707 | 0.728 | 0.716 | 0.2 | 0.237 | 0.206 |
| | Twitter | 0.652 | 0.688 | 0.667 | 0.192 | 0.197 | 0.19 |
| | Instagram | 0.638 | 0.705 | 0.664 | 0.263 | 0.277 | 0.256 |
| **2. Naïve Bayes** | Facebook | 0.716 | 0.763 | 0.719 | 0.17 | 0.347 | 0.193 |
| | Twitter | 0.709 | 0.769 | 0.679 | 0.148 | 0.324 | 0.193 |
| | Instagram | 0.589 | 0.757 | 0.663 | 0.163 | 0.092 | 0.066 |
| **3. Support Vector Machine** | Facebook | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Twitter | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.169 |
| | Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Facebook | 0.591 | 0.769 | 0.668 | 0.21 | 0.335 | 0.239 |
| | Twitter | 0.591 | 0.769 | 0.668 | 0.226 | 0.254 | 0.215 |
| | Instagram | 0.591 | 0.769 | 0.668 | 0.287 | 0.289 | 0.242 |

Figure 4.5 show the results of all classes. With *Random forest* classifier *Instagram* achieved highest F-measure of 0.256. *Facebook* achieved F-measure of 0.206 and *Twitter* achieved the lowest F-measure of 0.19. *Facebook* and *Twitter* achieved same F-measure with *Naïve Bayes* i.e. 0.193. However, *Instagram* achieved lowest F-measure of 0.066. *Twitter* achieved F-measure of 0.169 while *Facebook* and *Instagram* have achieved the F-measure of 0.168 with *Support Vector Machine*. With *Decision Tree, Instagram* obtained highest F-measure of 0.242. *Facebook* achieved 0.239 F-measure and *Twitter* achieved the lowest F-measure of 0.215.

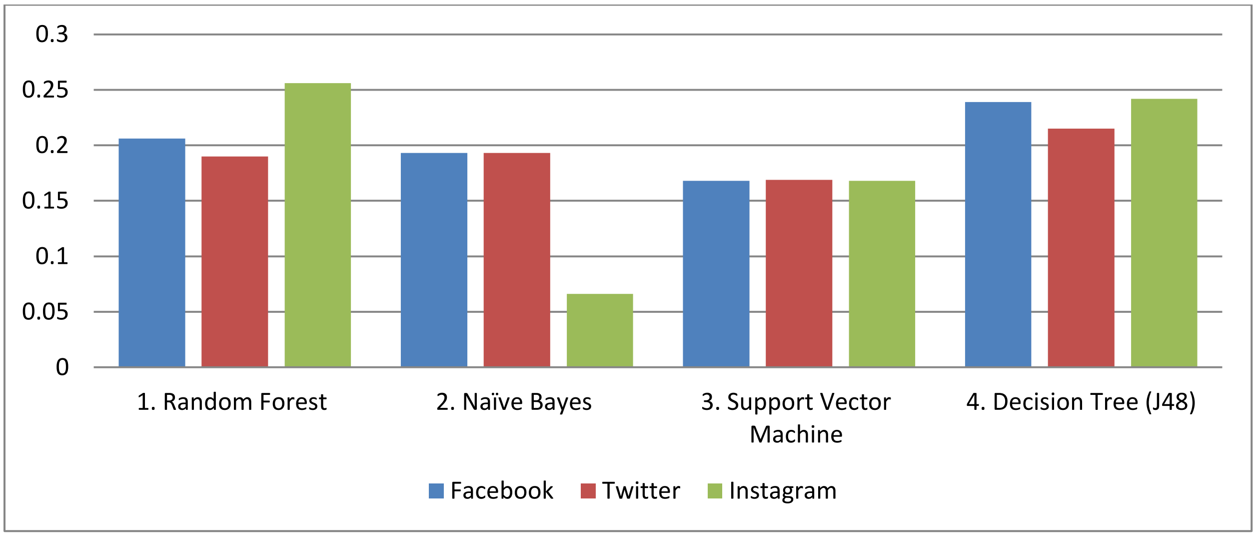


Figure 4.5 Actor 2- One Feature Analysis Class A to Class I

**Findings:** With *Random Forest Classifier, Instagram* obtained the best result by scoring F-measure of 0.256.

Figure 4.6 show the results of two classes. With *Random forest* classifier *Facebook* achieved the highest F-measure of 0.716 while *Twitter* achieved F-measure of 0.667 and *Facebook* achieved the lowest F-measure of 0.664. *Facebook* achieved the highest F-measure of 0.719 with *Naïve Bayes*. However, *Twitter* has achieved F-measure of 0.679 and *Instagram* achieved the lowest F-measure of 0.663. *Facebook*, *Twitter* and *Instagram* obtained the same result by scoring F-measure of 0.668. With *Decision Tree*, all these social media achieved the same F-measure of 0.668.

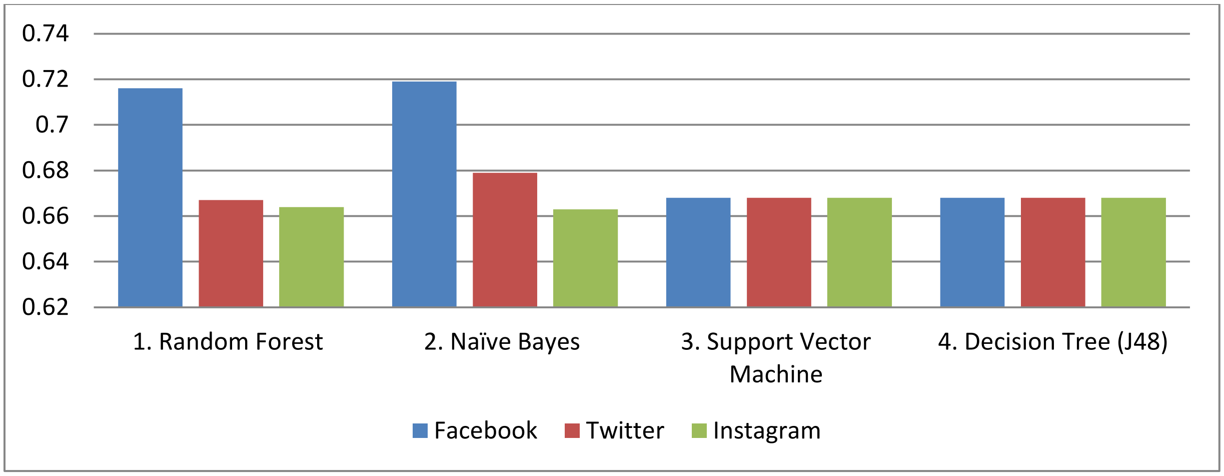


Figure 4.6 Actor 2- One Feature Analysis Successful and Un-successful Classes

**Findings:** *Facebook* obtained the best result by scoring F-measure of 0.719 with *Naïve Bayes Classifier*.

**(D) One Feature Analysis for Supporting Actor 3**

The results of one feature analysis for supporting actor 3 have been shown in this section.

Table 4.4: One Feature Analysis Supporting Actor 3 Social Media

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook | 0.65 | 0.659 | 0.654 | 0.2 | 0.237 | 0.206 |
| | Twitter | 0.624 | 0.647 | 0.635 | 0.192 | 0.197 | 0.19 |
| | Instagram | 0.65 | 0.723 | 0.675 | 0.263 | 0.277 | 0.256 |
| **2. Naïve Bayes** | Facebook | 0.591 | 0.769 | 0.668 | 0.17 | 0.347 | 0.193 |
| | Twitter | 0.774 | 0.78 | 0.704 | 0.148 | 0.324 | 0.193 |
| | Instagram | 0.591 | 0.769 | 0.668 | 0.163 | 0.092 | 0.066 |
| **3. Support Vector Machine** | Facebook | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Twitter | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.169 |
| | Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Facebook | 0.591 | 0.769 | 0.668 | 0.21 | 0.335 | 0.239 |
| | Twitter | 0.591 | 0.769 | 0.668 | 0.226 | 0.254 | 0.215 |
| | Instagram | 0.591 | 0.769 | 0.668 | 0.287 | 0.289 | 0.242 |

Results of all classes are shown Figure 4.7. With *Random forest* classifier *Instagram* achieved highest F-measure of 0.256. *Facebook* achieved F-measure of 0.206 and *Twitter* achieved the lowest F-measure of 0.19. *Facebook* and *Twitter* achieved same F-measure with *Naïve Bayes* i.e. 0.193. However, *Instagram* achieved lowest F-measure of 0.066. *Twitter* achieved F-measure of 0.169 while *Facebook* and *Instagram* have achieved the F-measure of 0.168 with *Support Vector Machine*. With *Decision Tree*, *Instagram* obtained highest F-measure of 0.242. *Facebook* achieved 0.239 F-measure and *Twitter* achieved the lowest F-measure of 0.215.

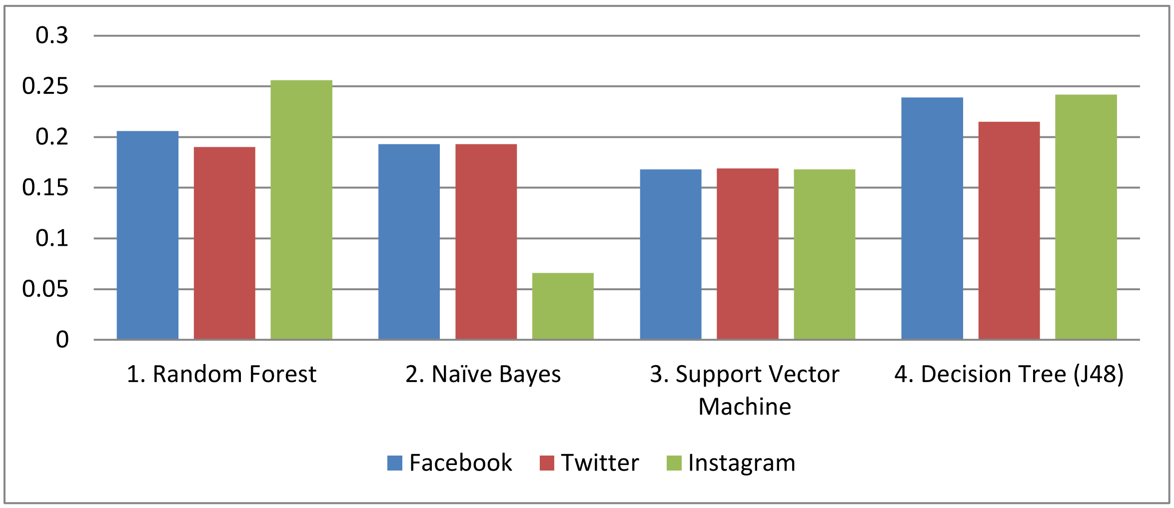


Figure 4.7: Actor 3- One Feature Analysis Class A to Class I

**Findings:** Again *Instagram* obtained the best result by scoring F-measure of 0.256 with *Random Forest Classifier*.

Figure 4.8 show the results of two classes. With *Random forest* classifier *Instagram* achieved the highest F-measure of 0.675 while *Facebook* achieved F-measure of 0.654 and *Twitter* achieved the lowest F-measure of 0.635. *Twitter* achieved the highest F-measure of 0.704 with *Naïve Bayes*. However, *Facebook* and *Instagram* achieved the F-measure of 0.668. *Facebook*, *Twitter* and *Instagram* obtained the same result by scoring F-measure of 0.668. With *Decision Tree*, all these social media achieved the same F-measure of 0.668.

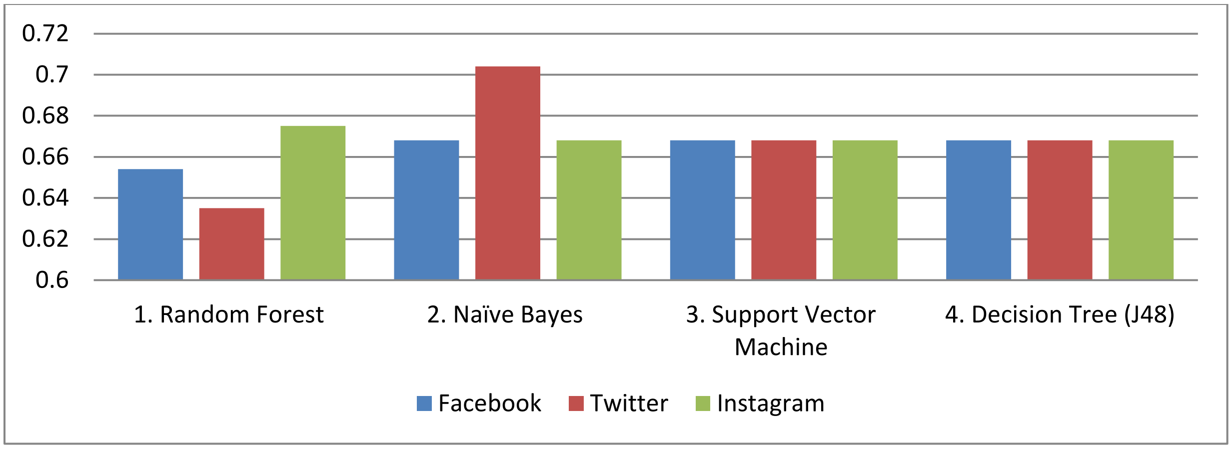


Figure 4.8: Actor 3- One Feature Analysis Successful and Un-successful Classes

**Findings:** Here with *Naïve Bayes Classifier*, *Twitter* obtained the best result by scoring F-measure of 0.704.

### 4.4.2 Two Feature Analyses

#### (A) Two Feature Analysis for Director

The results of two feature analysis for director have been shown in this section.

Table 4.4 Two Feature Analysis Director Social Media

| **Classifier** | | **Two Output Classes** | | | **All Classes** | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook+Twitter | 0.746 | 0.775 | 0.751 | 0.25 | 0.249 | 0.242 |
| | Facebook+Instagram | 0.719 | 0.746 | 0.729 | 0.268 | 0.243 | 0.244 |
| | Twitter+Instagram | 0.747 | 0.769 | 0.753 | 0.234 | 0.231 | 0.277 |
| **2. Naïve Bayes** | Facebook+Twitter | 0.738 | 0.775 | 0.7 | 0.186 | 0.324 | 0.201 |
| | Facebook+Instagram | 0.712 | 0.769 | 0.688 | 0.189 | 0.098 | 0.075 |
| | Twitter+Instagram | 0.637 | 0.306 | 0.269 | 0.123 | 0.087 | 0.067 |
| **3. Support Vector Machine** | Facebook+Twitter | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.169 |
| | Facebook+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.169 |
| **4. Decision Tree (J48)** | Facebook+Twitter | 0.591 | 0.769 | 0.668 | 0.23 | 0.249 | 0.227 |
| | Facebook+Instagram | 0.591 | 0.769 | 0.668 | 0.247 | 0.254 | 0.236 |
| | Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.317 | 0.306 | 0.289 |

Two classes' results are shown in the Figure 4.9. With *Random forest* classifier *Twitter+ Instagram* achieved highest F-measure of 0.753. *Facebook+ Twitter* achieved F-measure of 0.751 and *Facebook+ Instagram* achieved the lowest F-measure of 0.729. *Facebook+ Instagram* achieved highest F-measure with *Naïve Bayes* i.e. 0.688. However, *Twitter+ Instagram* achieved the F-measure of 0.269 and *Facebook+ Twitter* achieved the lowest F-measure i.e. 0.7. *Facebook+ Twitter*, *Facebook+ Instagram* and *Twitter+ Instagram* have achieved the same F-measure of 0.668 with *Support Vector Machine* and *Decision Tree.*

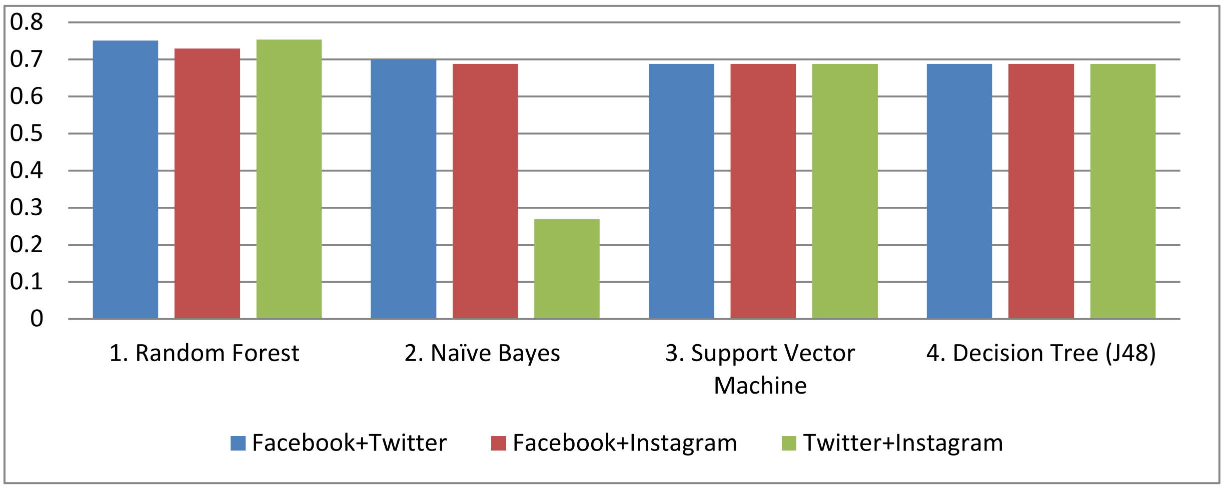


Figure 4.9 Director Two Feature Analyses with Successful and Un-successful Classes

**Findings:** When two features are used, the feature *Twitter+ Instagram* obtained the best result by scoring F-measure of 0.753 with *Random Forest Classifier*.

Figure 4.10 show the all classes results. With *Random forest* classifier *Twitter+ Instagram* achieved the F-measure of 0.277. *Facebook+ Instagram* achieved F-measure of 0.244 and *Facebook+ Twitter* achieved the lowest F-measure of 0.242. *Facebook+ Twitter* achieved highest F-measure with *Naïve Bayes* i.e. 0.201. However, *Facebook+ Instagram* achieved the F-measure of 0.075 and *Twitter+ Instagram* achieved the lowest F-measure i.e. 0.067. *Facebook+ Twitter* and *Twitter+ Instagram* have achieved the same F-measure of 0.169 with *Support Vector Machine*. *Facebook+ Instagram* achieved the F-measure of 0.168. With *Decision Tree, Twitter+ Instagram* obtained the best result by scoring F-measure of 0.289. *Facebook+ Instagram* achieved the F-measure of 0.236 and *Facebook+ Twitter* achieved 0.227.

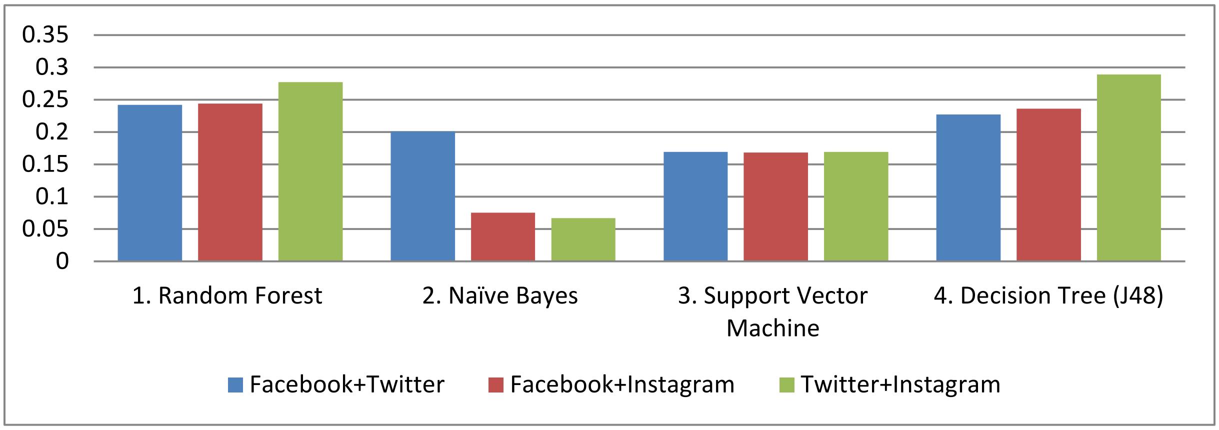


Figure 4.10 Director Two Feature Analyses with Class A to Class I

**Findings:** *Twitter+ Instagram* obtained the highest F-measure of 0.289 with *Decision Tree*.

**(B) Two Feature Analysis for Lead Actor 1**

The results of two feature analysis for lead actor 1 have been shown in this section.

Table 4.5: Two Feature Analysis Lead Actor Social Media

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook+Twitter | 0.651 | 0.711 | 0.673 | 0.198 | 0.214 | 0.199 |
| | Facebook+Instagram | 0.661 | 0.734 | 0.682 | 0.192 | 0.208 | 0.197 |
| | Twitter+Instagram | 0.7 | 0.74 | 0.713 | 0.24 | 0.225 | 0.229 |
| **2. Naïve Bayes** | Facebook+Twitter | 0.706 | 0.769 | 0.679 | 0.148 | 0.324 | 0.184 |
| | Facebook+Instagram | 0.59 | 0.763 | 0.665 | 0.237 | 0.335 | 0.227 |
| | Twitter+Instagram | 0.669 | 0.763 | 0.676 | 0.189 | 0.335 | 0.194 |
| **3. Support Vector Machine** | Facebook+Twitter | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Facebook+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Facebook+Twitter | 0.591 | 0.769 | 0.668 | 0.221 | 0.272 | 0.226 |
| | Facebook+Instagram | 0.591 | 0.769 | 0.668 | 0.209 | 0.277 | 0.238 |
| | Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.233 | 0.26 | 0.227 |

Two classes' results are shown in the Figure 4.11. With *Random forest* classifier *Twitter+ Instagram* achieved highest F-measure of 0.713. *Facebook+ Instagram* achieved F-measure of 0.682 and *Facebook+ Twitter* achieved the lowest F-measure of 0.673. *Facebook+ Twitter* achieved highest F-measure with *Naïve Bayes* i.e. 0.679. However, *Twitter+ Instagram* achieved the F-measure of 0.676 and *Facebook+ Instagram* achieved the lowest F-measure i.e. 0.665. *Facebook+ Twitter*, *Facebook+ Instagram* and *Twitter+ Instagram* have achieved the same F-measure of 0.668 with *Support Vector Machine*. With *Decision Tree*, again these features achieved the same F-measure of 0.668.

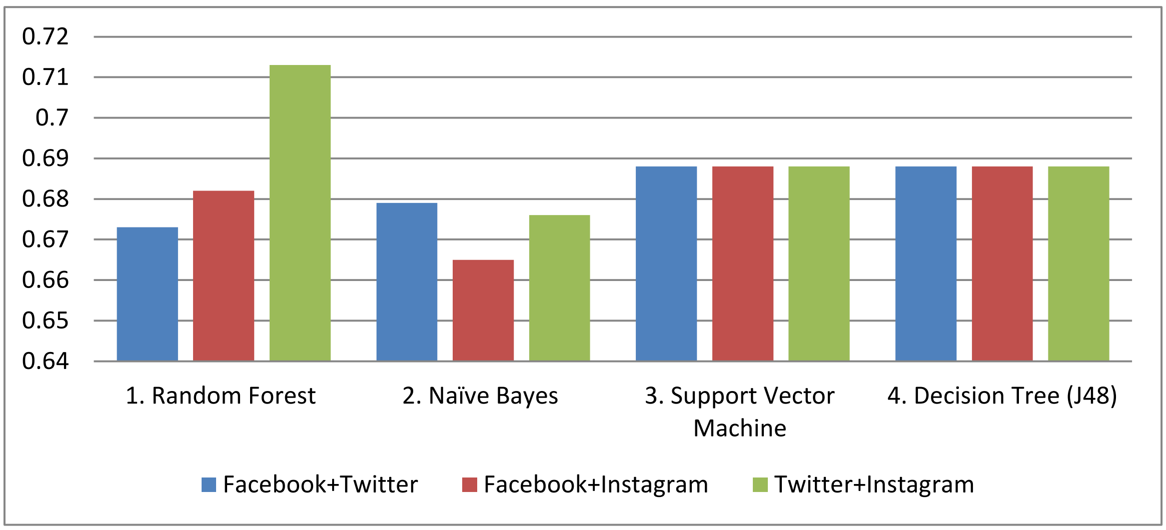


Figure 4.11 Actor one- Two Feature Analysis with Successful and Un-successful Classes

**Findings:** With *Random Forest*, *Twitter+ Instagram* obtained the best result by scoring F-measure of 0.713.

All classes' results are shown in the Figure 4.12. With *Random forest* classifier *Twitter+ Instagram* achieved highest F-measure of 0.229. *Facebook+ Twitter* achieved F-measure of 0.199 and *Facebook+ Instagram* achieved the F-measure of 0.197. *Facebook+ Instagram* achieved highest F-measure with *Naïve Bayes* i.e. 0.227. However, *Twitter+ Instagram* achieved the F-measure of 0.194 and *Facebook+ Twitter* achieved the lowest F-measure i.e. 0.184. *Facebook+ Twitter*, *Facebook+ Instagram* and *Twitter+ Instagram* have achieved the same F-measure of 0.168 with *Support Vector Machine*. With *Decision*

*Tree, Facebook+ Instagram* achieved the highest F-measure of 0.238. While *Twitter+ Instagram* achieved the F-measure of 0.227 and *Facebook+ Twitter* achieved F-measure of 0.226.

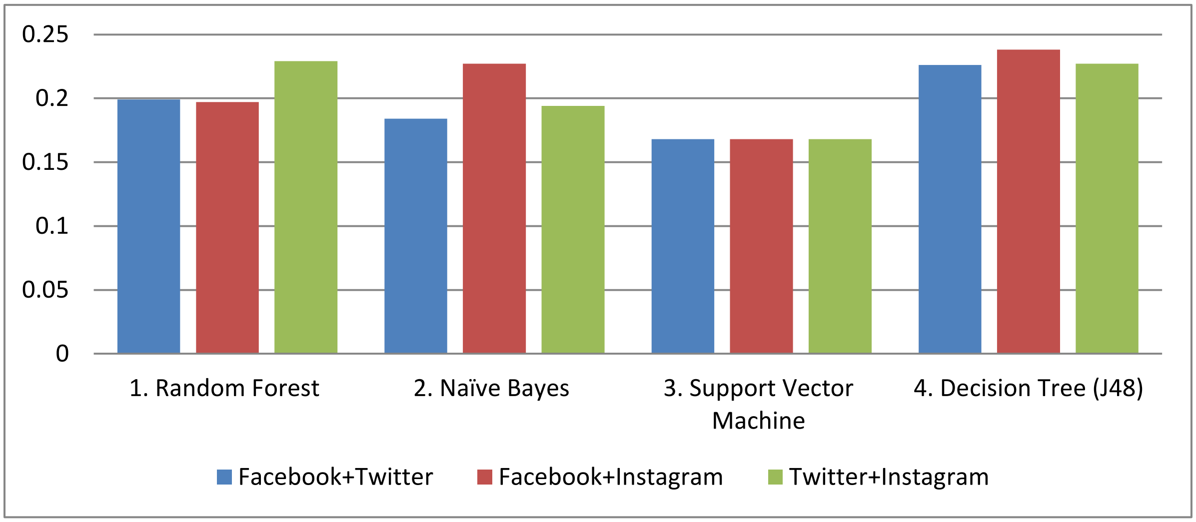


Figure 4.12 Actor one- Two Feature Analysis with Class A to Class I

**Findings:** Highest F-measure of 0.238 with *Decision Tree* has obtained by *Facebook+ Instagram.*

**(C) Two Feature Analysis for Supporting Actor 2**

The results of two feature analysis for supporting actor 2 have been shown in this section.

Table 4.6: Two Feature Analysis Supporting Actor 2 Social Media

| **Classifier** | | **Two Output Classes** | | | **All Classes** | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook+Twitter | 0.701 | 0.751 | 0.711 | 0.247 | 0.26 | 0.245 |
| | Facebook+Instagram | 0.667 | 0.723 | 0.686 | 0.23 | 0.214 | 0.208 |
| | Twitter+Instagram | 0.677 | 0.717 | 0.692 | 0.223 | 0.202 | 0.199 |
| **2. Naïve Bayes** | Facebook+Twitter | 0.716 | 0.763 | 0.719 | 0.165 | 0.11 | 0.095 |
| | Facebook+Instagram | 0.716 | 0.763 | 0.719 | 0.145 | 0.162 | 0.083 |
| | Twitter+Instagram | 0.649 | 0.757 | 0.672 | 0.124 | 0.092 | 0.055 |
| **3. Support Vector Machine** | Facebook+Twitter | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Facebook+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Facebook+Twitter | 0.591 | 0.769 | 0.668 | 0.286 | 0.254 | 0.234 |
| | Facebook+Instagram | 0.717 | 0.769 | 0.704 | 0.245 | 0.249 | 0.228 |
| | Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.21 | 0.26 | 0.227 |

Figure 4.13 show two classes results. With *Random forest* classifier *Facebook+ Twitter* achieved highest F-measure of 0.711. *Twitter+ Instagram* achieved F-measure of 0.692 and *Facebook+ Instagram* achieved the lowest F-measure of 0.686. *Facebook+ Twitter* and *Facebook+ Instagram* achieved highest F-measure with *Naïve Bayes* i.e. 0.719. However, *Twitter+ Instagram* achieved the F-measure of 0.672. *Facebook+ Twitter, Facebook+ Instagram* and *Twitter+ Instagram* have achieved the same F-measure of 0.668 with *Support Vector Machine*. With *Decision Tree, Facebook+ Instagram* achieved the F-measure of 0.704 while *Facebook+ Twitter* and *Twitter+ Instagram* have achieved the same F-measure of 0.668.

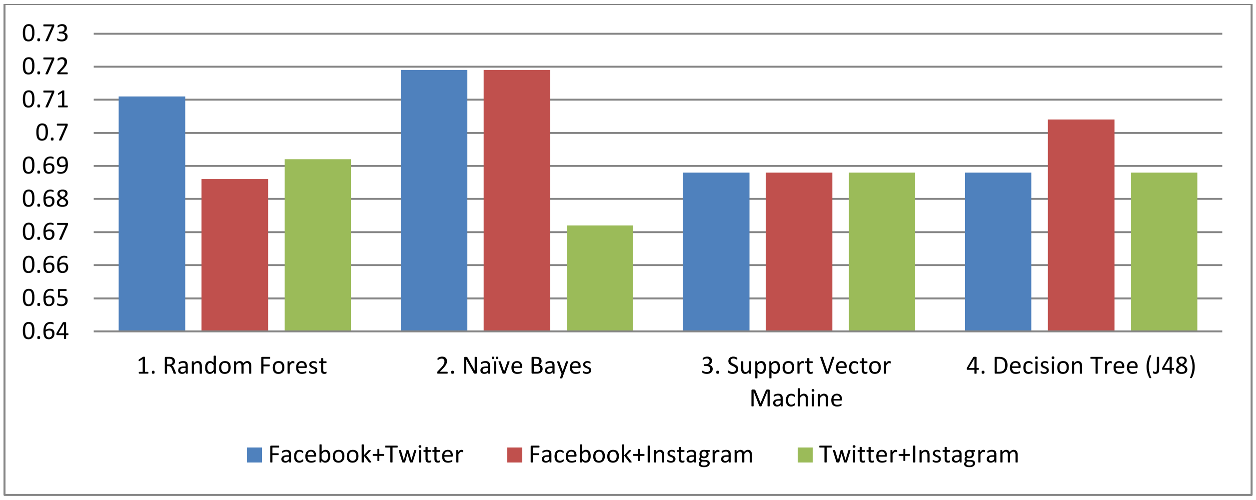


Figure 4.13 Actor 2- Two Feature Analysis with Successful and Un-successful Classes

**Findings:** *Facebook+ Instagram* and *Facebook+ Twitter* obtained the best result by scoring F-measure of 0.719 with *Naïve Bayes Classifier*.

All classes' results are shown in the Figure 4.14. With *Random forest* classifier *Facebook+ Twitter* obtained the best result by scoring F-measure of 0.245. *Facebook+ Instagram* achieved F-measure of 0.208 and *Twitter+ Instagram* achieved the F-measure of 0.199. *Facebook+ Twitter* achieved high F-measure with *Naïve Bayes* i.e. 0.095. However, *Facebook+ Instagram* achieved the F-measure of 0.083 and *Twitter+ Instagram* achieved the lowest F-measure i.e. 0.055. *Facebook+ Twitter, Facebook+ Instagram* and *Twitter+ Instagram* have achieved the same F-measure of 0.168 with *Support Vector Machine*. With Decision Tree, *Facebook+ Twitter* achieved the high F-measure of 0.234. While *Facebook+ Instagram* achieved the F-measure of 0.228 and *Twitter+ Instagram* achieved F-measure of 0.227.

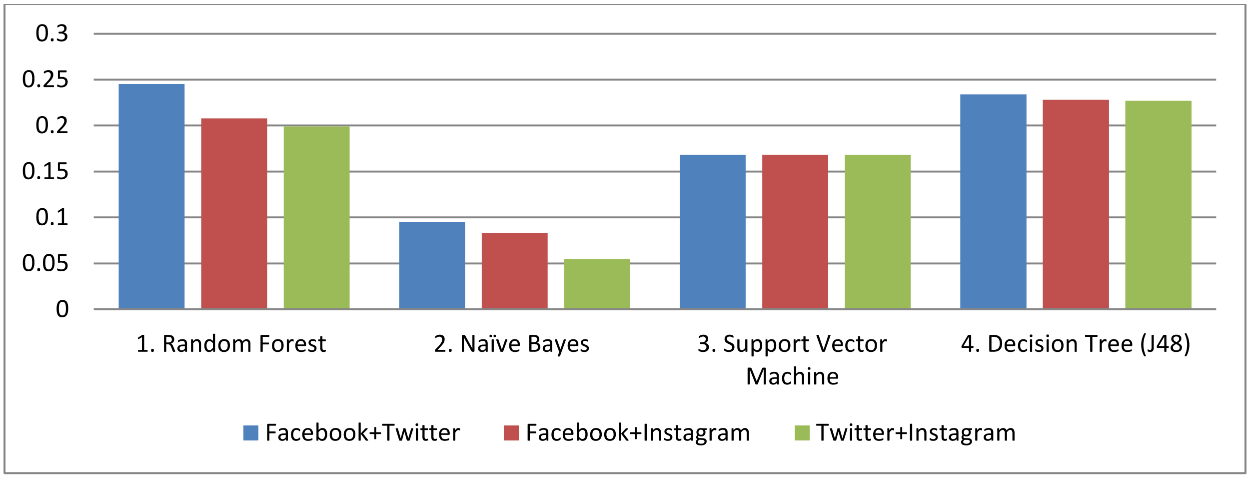


Figure 4.14 Actor 2- Two Feature Analysis with Class A to Class I

**Findings:** When two features are used, highest F-measure has obtained by feature *Facebook+ Twitter* with *Random Forest Classifier* i.e., 0.245.

**(D) Two Feature Analysis for Supporting Actor 3**

The results of two feature analysis for supporting actor 3 have been shown in this section.

Table 4.7: Two Feature Analysis Supporting Actor 3 Social Media

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook+Twitter | 0.659 | 0.723 | 0.681 | 0.178 | 0.179 | 0.174 |
| | Facebook+Instagram | 0.691 | 0.728 | 0.704 | 0.171 | 0.185 | 0.176 |
| | Twitter+Instagram | 0.69 | 0.734 | 0.704 | 0.237 | 0.214 | 0.217 |
| **2. Naïve Bayes** | Facebook+Twitter | 0.774 | 0.78 | 0.704 | 0.166 | 0.289 | 0.204 |
| | Facebook+Instagram | 0.589 | 0.757 | 0.663 | 0.04 | 0.15 | 0.053 |
| | Twitter+Instagram | 0.789 | 0.786 | 0.715 | 0.078 | 0.156 | 0.063 |
| **3. Support Vector Machine** | Facebook+Twitter | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Facebook+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Facebook+Twitter | 0.591 | 0.769 | 0.668 | 0.237 | 0.243 | 0.227 |
| | Facebook+Instagram | 0.591 | 0.769 | 0.668 | 0.211 | 0.26 | 0.224 |
| | Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.213 | 0.208 | 0.202 |

Two classes' results are shown in the Figure 4.15. With *Random forest* classifier *Facebook+ Instagram* and *Twitter+ Instagram* achieved F-measure of 0.704 while *Facebook+ Twitter* achieved the F-measure of 0.681. *Twitter+ Instagram* achieved highest F-measure with *Naïve Bayes* i.e. 0.715. However, *Facebook+ Twitter* achieved F-measure of 0.704 and *Facebook+ Instagram* achieved the F-measure of 0.663. *Facebook+ Twitter*, *Facebook+ Instagram* and *Twitter+ Instagram* have achieved the same F-measure of 0.668 with *Support Vector Machine*. With *Decision Tree, Facebook+ Instagram* achieved the F-measure of 0.704 while *Facebook+ Twitter* and *Twitter+ Instagram* have achieved the same F-measure of 0.668**.**

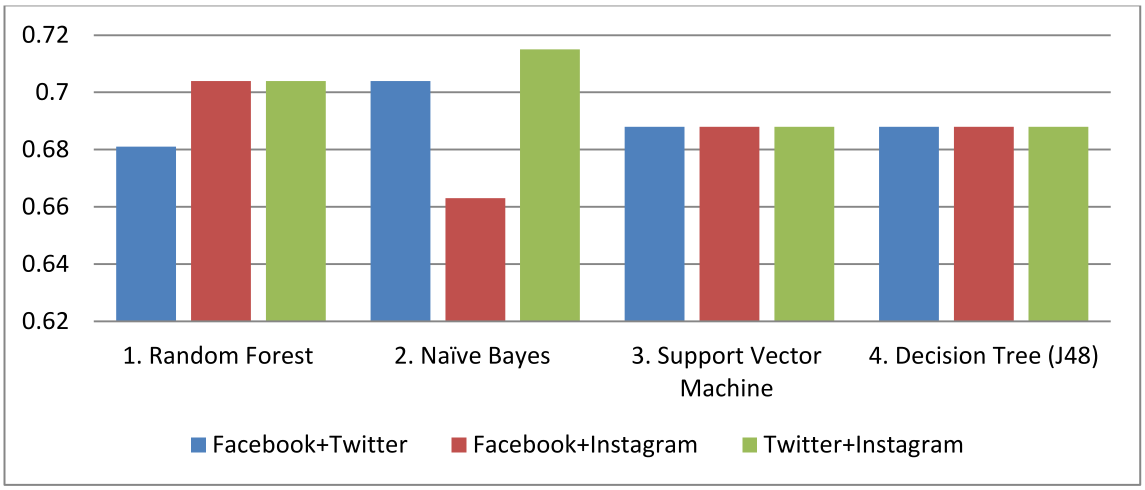


Figure 4.15 Actor 3- Two Feature Analysis with Successful and Un-successful Classes

**Findings:** *Twitter+ Instagram* obtained the highest F-measure of 0.715 with *Naïve Bayes Classifier*.

Figure 4.16 show results of all classes. With *Random forest* classifier *Twitter+ Instagram* obtained the best result by scoring F-measure of 0.217. *Facebook+ Instagram* achieved F-measure of 0.176 and *Facebook+ Twitter* achieved the F-measure of 0.174. *Facebook+ Twitter* achieved high F-measure with *Naïve Bayes* i.e. 0.204. However, *Twitter+ Instagram* achieved the F-measure of 0.063 and *Facebook+ Instagram* achieved the lowest F-measure i.e. 0.053. *Facebook+ Twitter*, *Facebook+ Instagram* and *Twitter+ Instagram* have achieved the same F-measure of 0.168 with *Support Vector Machine*. With Decision Tree, *Facebook+ Twitter* achieved the high F-measure of 0.227. While

*Facebook+ Instagram* achieved the F-measure of 0.224 and *Twitter+ Instagram* achieved F-measure of 0.202.

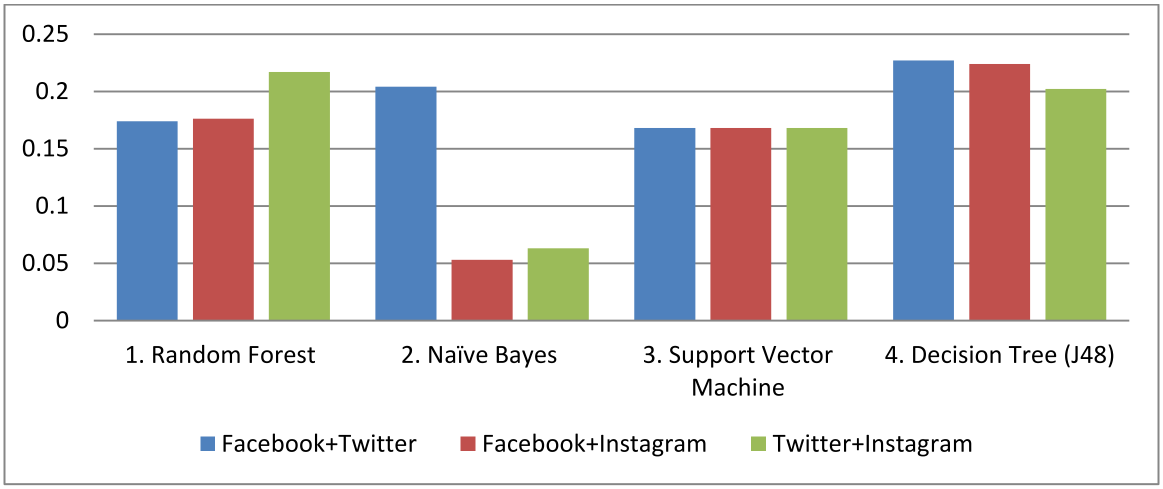


Figure 4.16 Actor 3- Two Feature Analysis with Class A to Class I

**Findings:** With *Decision Tree*, *Facebook+ Twitter* obtained the best result of 0.227.

### 4.4.3 Three Feature Analyses

**(A) Three Feature Analysis for Director**

The results of three feature analysis for director have been shown in this section.

Table 4.8: Three Feature Analysis Director Social Media

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook+Twitter+Instagram | 0.784 | 0.803 | 0.781 | 0.279 | 0.266 | 0.257 |
| **2. Naïve Bayes** | Facebook+Twitter+Instagram | 0.712 | 0.769 | 0.688 | 0.18 | 0.087 | 0.073 |
| **3. Support Vector Machine** | Facebook+Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.169 |
| **4. Decision Tree (J48)** | Facebook+Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.216 | 0.214 | 0.208 |

Figure 4.17 show the results of two classes. With *Random forest* classifier *Facebook+Twitter+Instagram* achieved highest the F-measure of 0.781. With *Naïve Bayes*, *Support Vector Machine* and *Decision Tree* the F-measure of 0.668 is achieved.

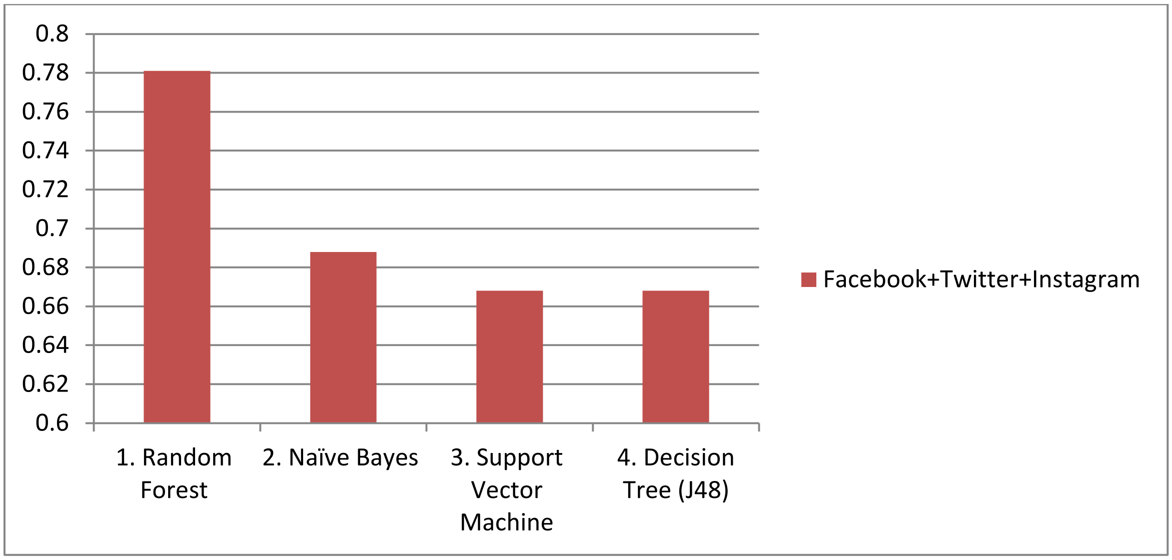


Figure 4.17 Director Three Feature Analyses with Successful and Un-successful Classes

**Findings:** F-measure of 0.781 has obtained by *Random Forest Classifier*.

All classes' results are shown in Figure 4.18. With *Random forest* classifier *Facebook+Twitter+Instagram* achieved highest the F-measure of 0.257. With *Naïve Bayes*, 0.073 F-measure is achieved. *Support Vector Machine* achieved 0.169 F-measure and with *Decision Tree* the F-measure of 0.208 is achieved.

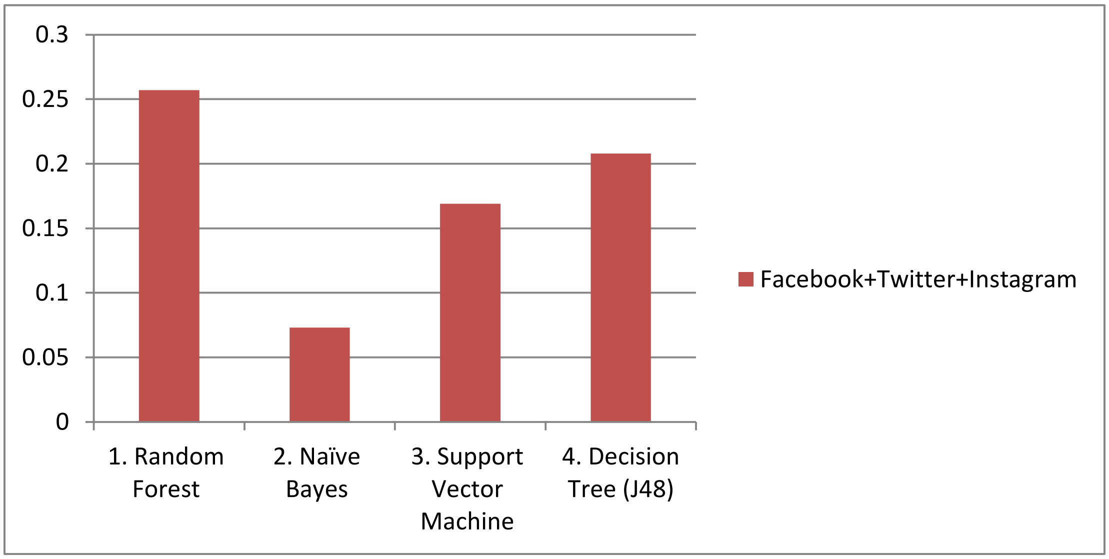


Figure 4.18 Director Three Feature Analysis with Class A to I

**Findings:** *Random Forest Classifier* obtained the best result by scoring F-measure of 0.257.

### (B) Three Feature Analysis for Lead Actor 1

The results of three feature analysis for lead actor 1 have been shown in this section.

Table 4.9: Three Feature Analysis Lead Actor Social Media

| **Classifier** | | **Two Output Classes** | | | **All Classes** | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook+Twitter+Instagram | 0.682 | 0.746 | 0.696 | 0.229 | 0.243 | 0.224 |
| **2. Naïve Bayes** | Facebook+Twitter+Instagram | 0.669 | 0.763 | 0.676 | 0.196 | 0.324 | 0.209 |
| **3. Support Vector Machine** | Facebook+Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Facebook+Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.209 | 0.249 | 0.222 |

Two classes' results are shown in Figure 4.19. With *Random forest* classifier *Facebook+Twitter+Instagram* achieved highest the F-measure of 0.696. With *Naïve Bayes* F-measure of 0.676 is achieved. With *Support Vector Machine* and *Decision Tree* the F-measure of 0.668 is achieved.

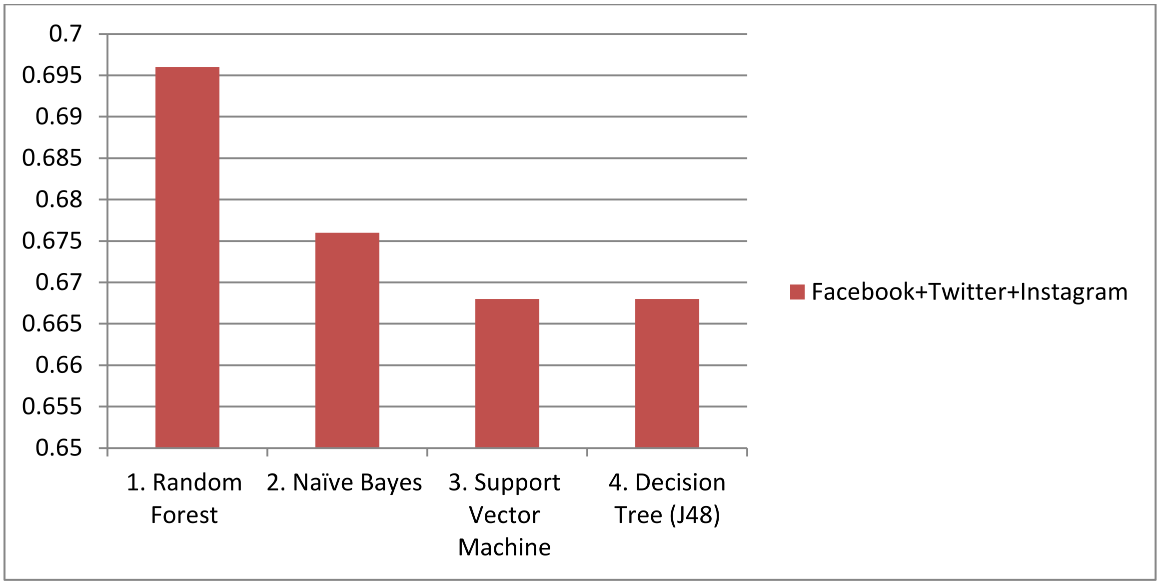


Figure 4.19 Actor one- Three Feature Analysis with Successful and Un-successful Classes

**Findings:** Highest F-measure of 0.696 has obtained by *Random Forest Classifier*.

In Figure 4.20 results of two classes are shown. With *Random forest* classifier *Facebook+Twitter+Instagram* achieved highest the F-measure of 0.224. With *Naïve Bayes* 0.209 F-measure is achieved. With *Support Vector Machine* the F-measure of

0.168 is achieved and with *Decision Tree* 0.222 F-measure is obtained.

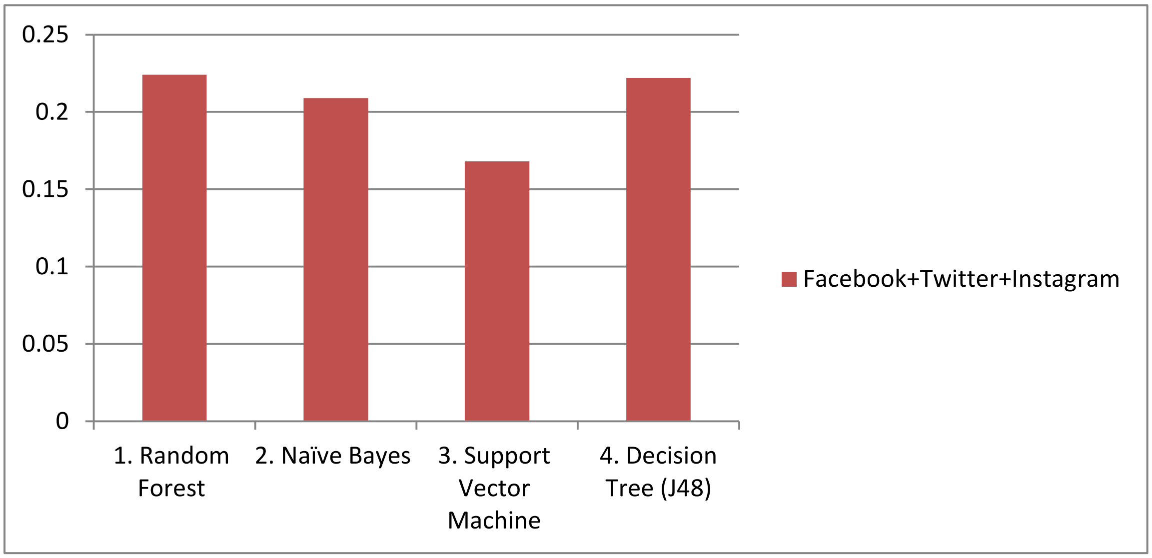


Figure 4.20 Actor one- Three Feature Analysis with Class A to Class I

**Findings:** *Random Forest Classifier* obtained the best result by scoring F-measure of 0.224.

**(C) Three Feature Analysis for Supporting Actor 2**

The results of three feature analysis for supporting actor 2 have been shown in this section.

Table 4.10: Three Feature Analysis Supporting Actor 2 Social Media

| **Classifier** | | **Two Output Classes** | | | **All Classes** | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Facebook+Twitter+Instagram | 0.716 | 0.763 | 0.719 | 0.262 | 0.22 | 0.206 |
| **2. Naïve Bayes** | Facebook+Twitter+Instagram | 0.716 | 0.763 | 0.719 | 0.071 | 0.162 | 0.076 |
| **3. Support Vector Machine** | Facebook+Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Facebook+Twitter+Instagram | 0.717 | 0.769 | 0.704 | 0.248 | 0.289 | 0.263 |

Figure 4.21 show the results of two classes. With *Random forest* and *Naïve Bayes* classifiers *Facebook+Twitter+Instagram* achieved highest the F-measure of 0.719. While *Support Vector Machine* achieved 0.668 and *Decision Tree* achieved 0.704 F-measure.

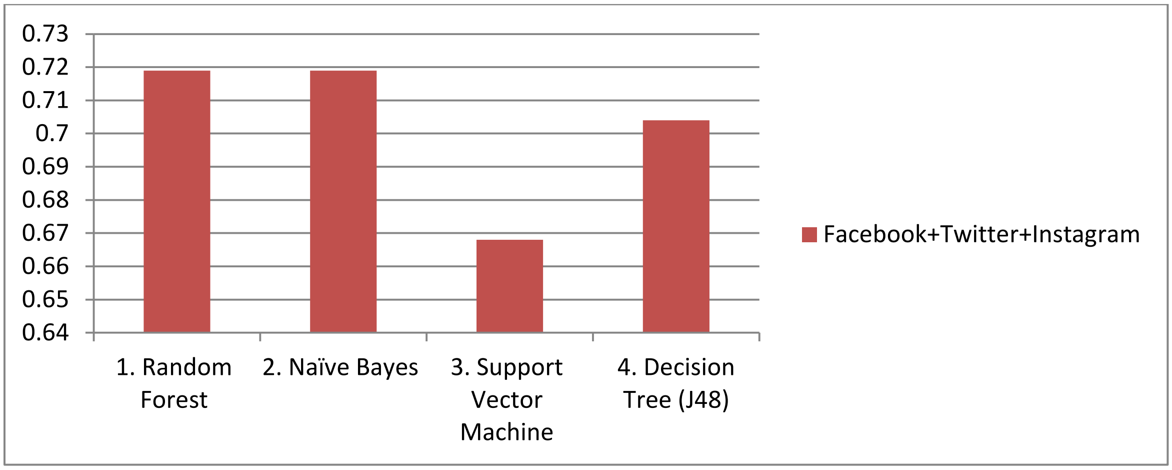


Figure 4.21 Actor 2- Three Feature Analysis with Successful and Un-successful Classes

**Findings:** *Random Forest* and *Naïve Bayes Classifier* obtained the highest F-measure of 0.719.

Figure 4.22 show the results of two classes. With *Random forest* classifier *Facebook+Twitter+Instagram* achieved the F-measure of 0.206. With *Naïve Bayes* lowest F-measure of 0.076 is achieved. With *Support Vector Machine* 0.016 F-measure is achieved. *Decision Tree* obtained the highest F-measure of 0.263.

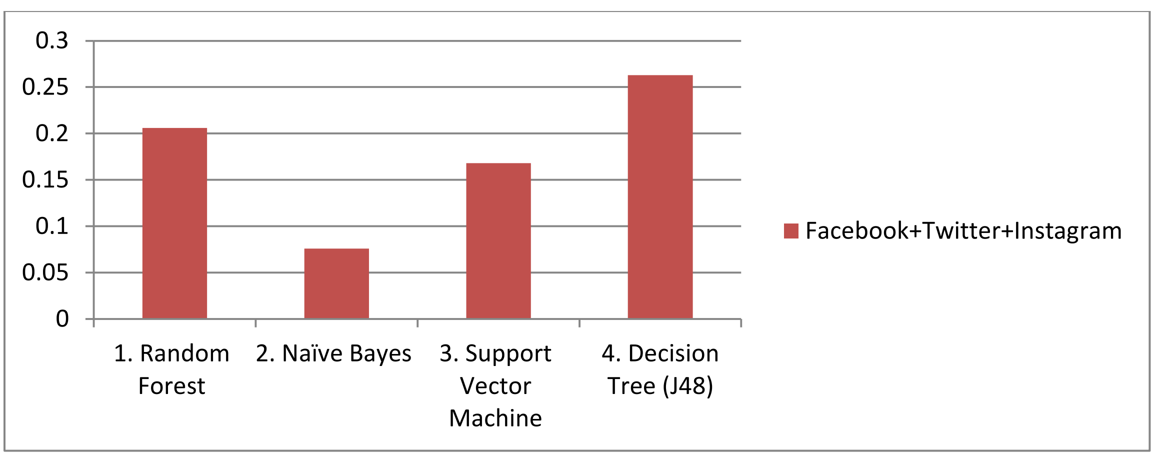


Figure 4.22 Actor 2- Three Feature Analysis with Class A to Class I

**Findings:** Highest F-measure has obtained by *Decision Tree* i.e. 0.263.

### (D) Three Feature Analysis for Supporting Actor 3

The results of three feature analysis for supporting actor 3 have been shown in this section.

Table 4.11: Three Feature Analysis Supporting Actor 3 Social Media

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | Feature Name | Precision | Recall | F-Measure | Precision | Recall | F-Measure |
| 1. Random Forest | Facebook+Twitter+Instagram | 0.69 | 0.751 | 0.7 | 0.187 | 0.202 | 0.187 |
| 2. Naïve Bayes | Facebook+Twitter+Instagram | 0.774 | 0.78 | 0.704 | 0.034 | 0.139 | 0.049 |
| 3. Support Vector Machine | Facebook+Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| 4. Decision Tree (J48) | Facebook+Twitter+Instagram | 0.591 | 0.769 | 0.668 | 0.19 | 0.179 | 0.18 |

Figure 4.23 show the results of two classes. With *Random forest* classifier *Facebook+Twitter+Instagram* achieved the F-measure of 0.7. With *Naïve Bayes* highest F-measure of 0.704 is achieved. With *Support Vector Machine* and *Decision Tree* the F-measure of 0.668 is achieved.

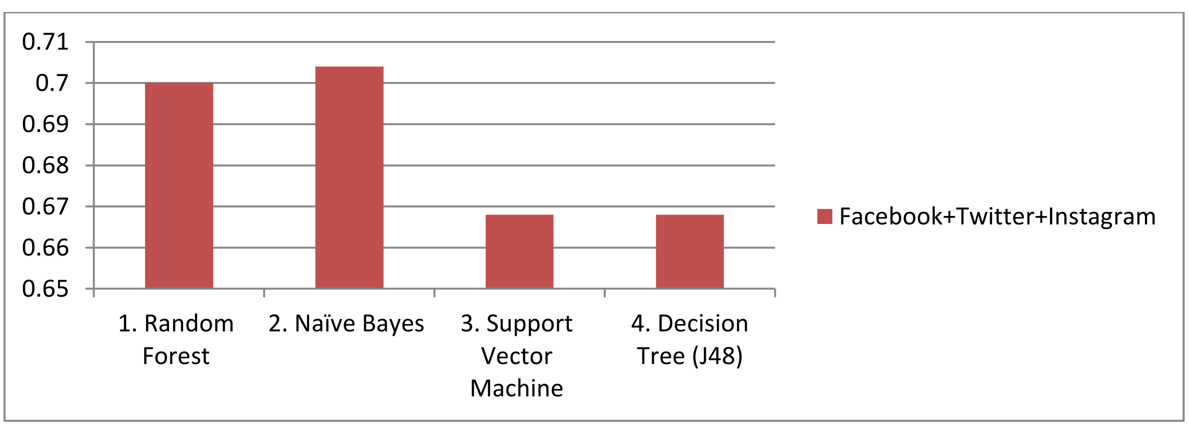

Figure 4.23 Actor 3- Three Feature Analysis with Successful and Un-successful Classes

**Findings:** Here *Naïve Bayes* has obtained the best result of 0.704.

In Figure 4.24 all classes' results are shown. With *Random forest* classifier *Facebook+Twitter+Instagram* achieved highest the F-measure of 0.187. With *Naïve Bayes* lowest F-measure of 0.049 is achieved. With *Support Vector Machine* 0.168 F-measure is achieved and with *Decision Tree* the F-measure of 0.18 is achieved.

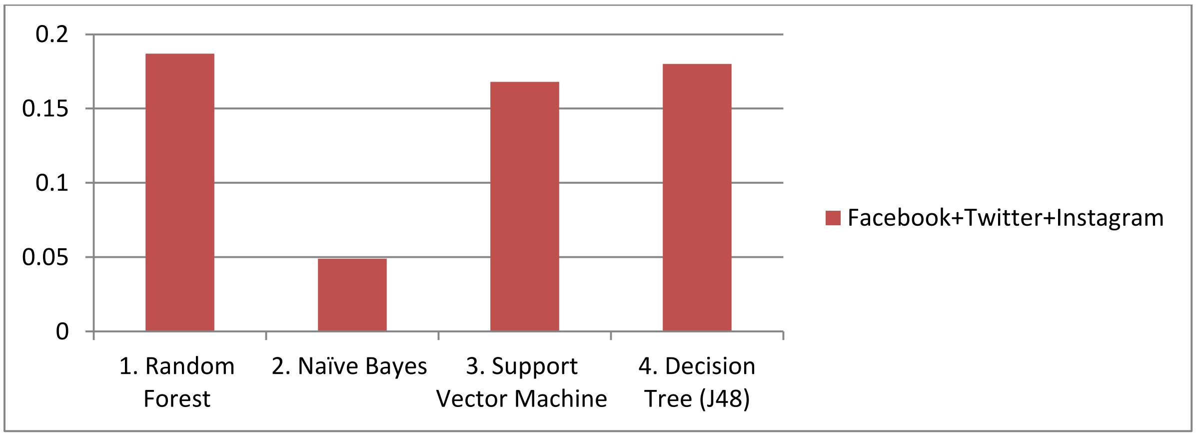


Figure 4.24 Actor 3- Three Feature Analysis with Class A to I

**Findings:** 0.187 F-measure has obtained by *Random Forest Classifier*.

| Feature Analysis | All Classes | | Two Class | |
|---|---|---|---|---|
| | **Feature Name** | **F1 Measure** | **Feature Name** | **F1 Measure** |
| **One Feature Analysis** | Instagram | 0.256 | Instagram | 0.733 |
| **Two Feature Analysis** | Twitter+ Instagram | 0.289 | Twitter+ Instagram | 0.753 |
| **Three Feature Analysis** | Facebook+ Twitter+ Instagram | 0.263 | Facebook+ Twitter+ Instagram | 0.781 |

Table 4.12: Results Conclusion

## 4.5 Awards Impact on Movie Success

In this section, the following research question has been evaluated:

**Does the following awards: Oscar Awards, Golden Globe Awards and Venice Awards won by directors, lead and supporting actress/actor matter in the success of a Hollywood movie?**

To answer this question, the same methodology has been adopted as was discussed to answer the research question number 1 in the Section 4.4.

### 4.5.1 One Feature Analysis

**(A) One Feature Analysis for Director**

The results of one feature analysis for director have been shown in this section.

Table 4.13 One Feature Analysis Director Awards

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar Award | 0.59 | 0.763 | 0.665 | 0.15 | 0.306 | 0.178 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.208 | 0.335 | 0.19 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.281 | 0.341 | 0.18 |
| **2. Naïve Bayes** | Oscar Award | 0.628 | 0.746 | 0.666 | 0.158 | 0.318 | 0.194 |
| | Golden Globe | 0.589 | 0.757 | 0.663 | 0.164 | 0.347 | 0.204 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.124 | 0.156 | 0.09 |
| **3. Support Vector Machine** | Oscar Award | 0.591 | 0.769 | 0.668 | 0.151 | 0.324 | 0.196 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.028 | 0.168 | 0.048 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Oscar Award | 0.591 | 0.769 | 0.668 | 0.172 | 0.312 | 0.187 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.198 | 0.341 | 0.181 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |

The complete results have been shown in the Table 4.12 and the results of all classes are shown in the Figure 4.25. This has been highlighted that how each of the evaluated feature performed to classify the review into all 9 distinct classes. Firstly we applied *Random forest* classifier on the

Feature set 1. If we observe the F-measure closely, *Golden Globe Award* performed the best by securing the F-measure of 0.19. Similarly following forms were able to achieve the F- measure of more than or equal 0.17 i.e. *Venice Award* and *Oscar Award.* However, *Oscar Award* obtained lowest F- measure of 0.17. With *Naïve Bayes* again *Golden Globe Award* has higher F-Measure and *Venice Award* has lowest F-measure. *Oscar Award* obtained the best result by scoring F-measure of 0.196 with *Support Vector Machine*. It is also revealed that using *Decision Tree, Oscar Award* has highest F-measure of 0.187. The *Golden Globe Award* has F-measure of 0.181 and *Venice Award* has 0.168. Overall, these results are not encouraging because, the classification was done into nine classes. The results of two classes: Successful and Un-successful classes are quite encouraging as shown in the Figure 4.26.

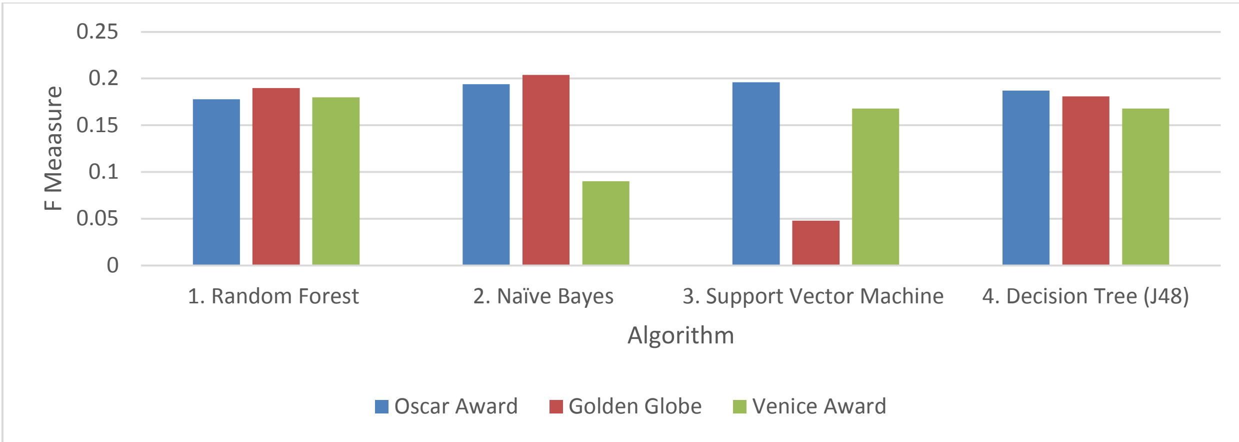


Figure 4.25 Director One Feature Analysis with Class A to Class I

**Findings:** The feature *Golden Globe* obtained the highest F-measure of 0.20 with *Naïve Bayes Classifier*.

The results of two classes are shown in the Figure 4.26. If we take a look at *Random Forest's* result, it is clear that *Golden Globe Award* and *Venice Award* scored equal F-measure which is 0.668 while *Oscar Award* scored lowest F-measure of 0.665. *Venice Award* scored high F-measure of 0.668 using *Naïve Bayes* classifier. *Oscar Award* obtained F-measure of 0.666 however; Golden Globe Award achieved lowest F-measure of 0.663. It has been seen that with *Support Vector Machine*, *Oscar Award*, *Golden Globe Award* and *Venice Award* obtained same F-measure of 0.668. Also, these awards achieved same F-measure of 0.668 with *Decision Tree*.

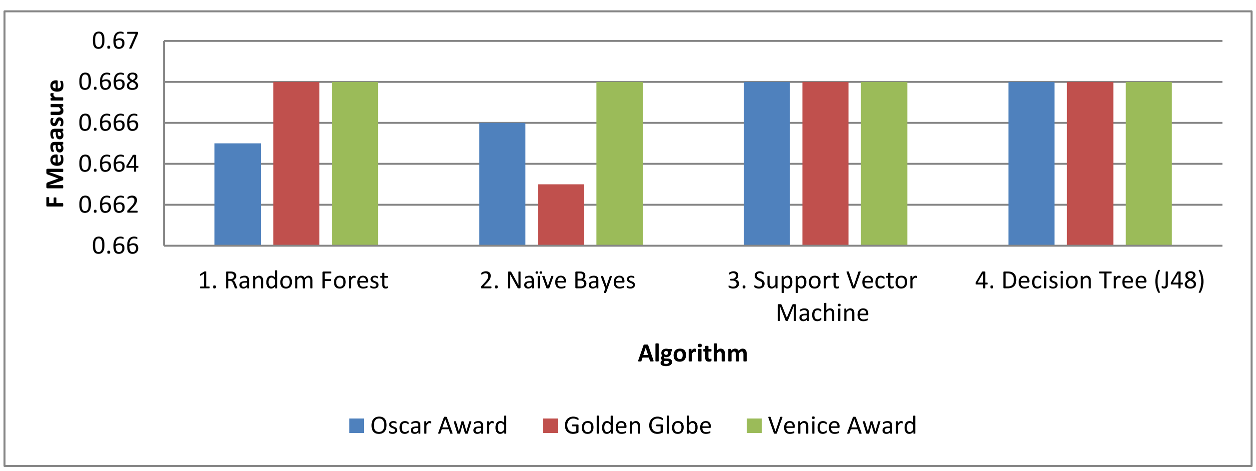


Figure 4.26 Director One Feature Analysis Successful and Un-successful Classes

**Findings:** When single feature is used, it has been seen that *Golden Globe* and *Venice* awards achieved high F-measure of 0.668 with *Random Forest*. With *Naïve Bayes*, *Venice Award* achieved F-measure of 0.688. Similarly *Oscar*, *Golden Globe* and *Venice Awards* achieved F-measure of 0.688 with *Support Vector Machine* and *Decision Tree respectively*.

**(B) One Feature Analysis for Lead Actor**

The results of one feature analysis for lead actor have been shown in this section.

Table 4.14 One Feature Analysis Lead Actor Awards

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar Award | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.169 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.199 | 0.347 | 0.192 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.169 |
| **2. Naïve Bayes** | Oscar Award | 0.826 | 0.775 | 0.682 | 0.113 | 0.335 | 0.169 |
| | Golden Globe | 0.751 | 0.775 | 0.691 | 0.2 | 0.341 | 0.183 |
| | Venice Award | 0.589 | 0.757 | 0.663 | 0.112 | 0.329 | 0.167 |
| **3. Support Vector Machine** | Oscar Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Oscar Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.198 | 0.341 | 0.181 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |

The results of all classes are shown in the Figure 4.27. Firstly we applied *Random forest* classifier and we observed that *Golden Globe Award* performed the best by securing the F-measure of 0.192. While *Venice Award* and *Oscar Award* were able to achieve the F-measure of 0.169. With *Naïve Bayes Golden Globe Award* has higher F-measure of 0.183 and *Venice Award* has lower F-measure of 0.167. While *Oscar Award* obtained F-measure of 0.169. With *Support Vector Machine Oscar Award*, *Golden Globe Award* and *Venice Award* obtained the same result by scoring F-measure of 0.168. It is also reveal that using *Decision Tree, Golden Globe Award* has highest F-measure of 0.181. The *Oscar Award* and *Venice Award* have same F-measure of 0.168.

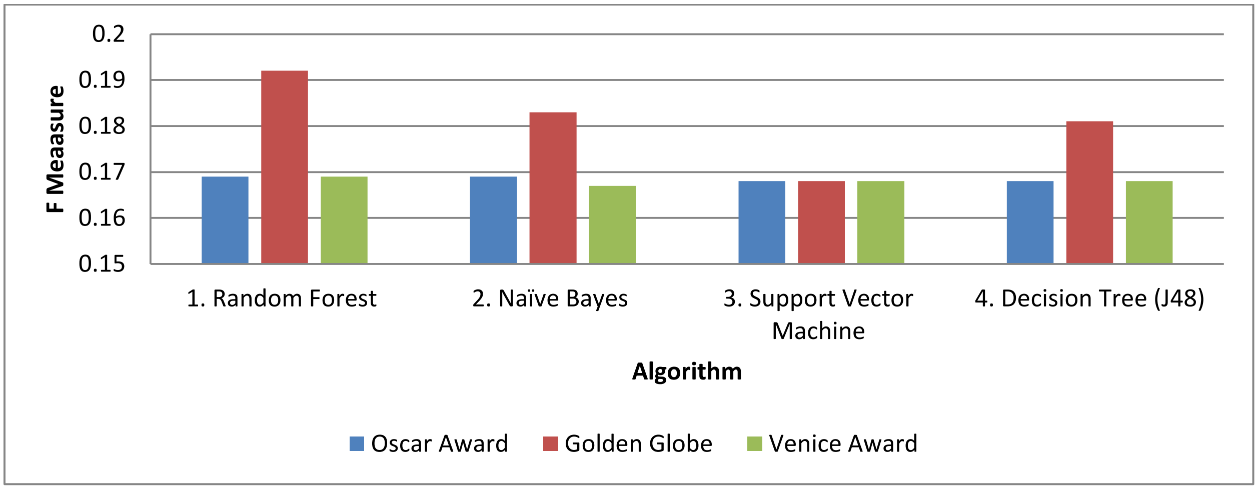


Figure 4.27 Actor One - One Feature Analysis Class A to Class I

**Findings:** *Golden Globe* obtained the highest F-measure of 0.192 with *Random Forest Classifier*.

Figure 4.28 show two classes results. After applying *Random forest* classifier, it has been observed that *Oscar Award, Golden Globe Award* and *Venice Award* achieved the same result by scoring the F-measure of 0.668. With *Naïve Bayes classifier, Golden Globe Award* has higher F-measure of 0.691 and *Venice Award* has lower F-measure of 0.663. While *Oscar Award* obtained F-measure of 0.682. *Oscar Award, Golden Globe Award* and *Venice Award* obtained the same result by scoring F-measure of 0.668 with *Support Vector Machine*. Also with *Decision Tree,* these awards obtained same F-measure of 0.668.

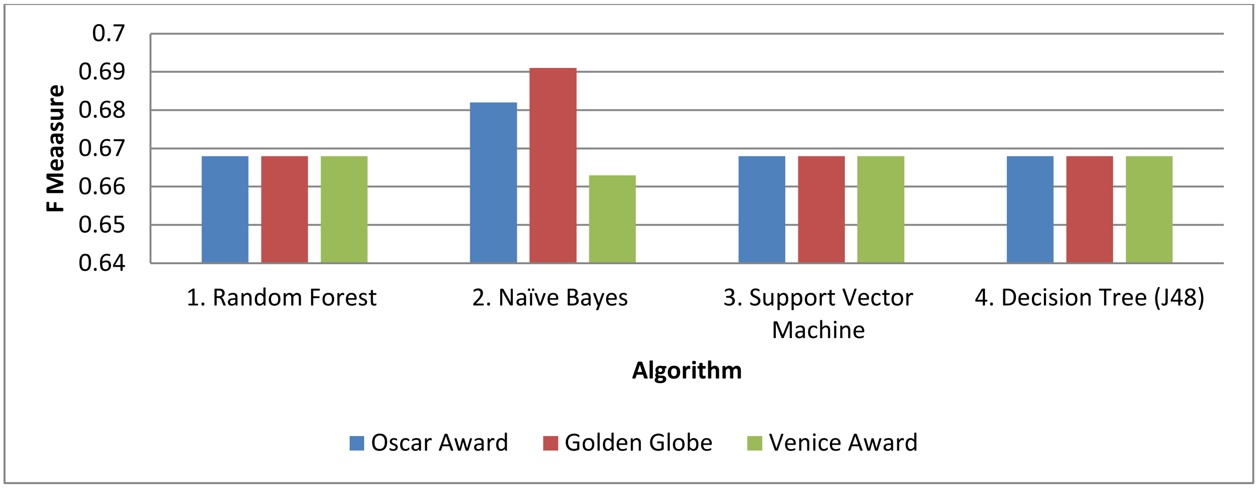


Figure 4.28 Actor One - One Feature Analysis Successful and Un-successful Classes

**Findings:** With *Naïve Bayes Classifier*, *Golden Globe* obtained the best result of 0.69.

### (C) One Feature Analysis for Supporting Actor 2

The results of one feature analysis for supporting actor 2 have been shown in this section.

Table 4.15: One Feature Analysis Supporting Actor 2 Award

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar Award | 0.591 | 0.769 | 0.668 | 0.2 | 0.329 | 0.181 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.2 | 0.341 | 0.183 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.163 | 0.341 | 0.181 |
| **2. Naïve Bayes** | Oscar Award | 0.656 | 0.74 | 0.679 | 0.153 | 0.335 | 0.188 |
| | Golden Globe | 0.672 | 0.751 | 0.686 | 0.136 | 0.324 | 0.175 |
| | Venice Award | 0.59 | 0.763 | 0.665 | 0.113 | 0.335 | 0.169 |
| **3. Support Vector Machine** | Oscar Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Oscar Award | 0.591 | 0.769 | 0.668 | 0.2 | 0.329 | 0.181 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |

All classes' results are shown in the Figure 4.29. With *Random forest* classifier *Golden Globe Award* achieved highest F-measure of 0.183. While *Oscar Award* and *Venice Award* achieved the F-measure of 0.181. Oscar Award achieved highest F-measure with *Naïve Bayes* i.e. 0.188. However, *Golden Globe Award* achieved 0.175 F-measure and *Venice Award* has lower F-measure of 0.169. *Oscar Award*, *Golden Globe Award* and *Venice Award* obtained the same result by scoring F-measure of 0.168 with *Support Vector Machine*. With *Decision Tree, Oscar Awards* obtained highest F-measure of 0.181 and *Golden Globe Award* and *Venice Award* achieved F-measure of 0.168.

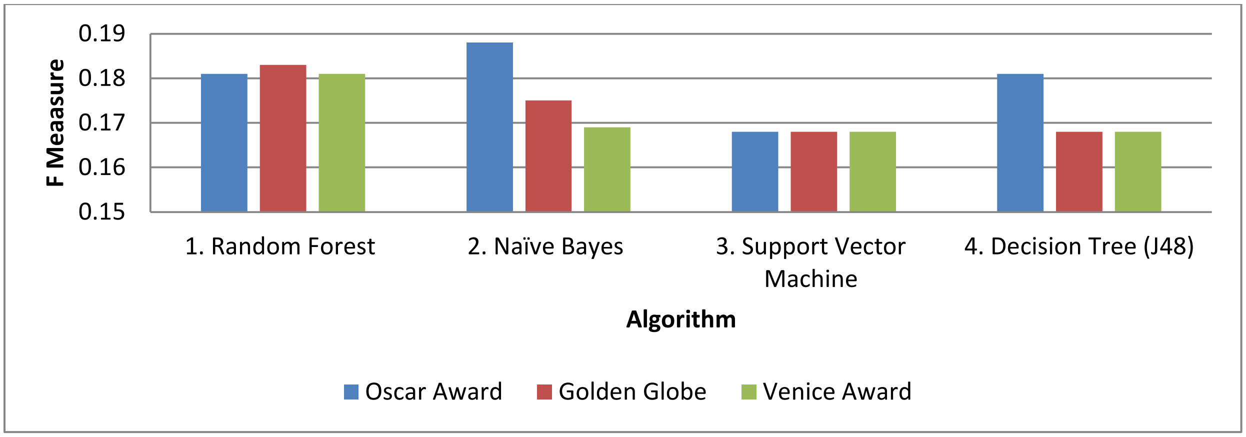


Figure 4.29 Actor 2 - One Feature Analysis Class A to Class I

**Findings:** F-measure of 0.188 has been obtained by *Oscar Award* with *Naïve Bayes Classifier*.

Two classes (Successful movie class and Un-successful movie class) results are shown in the Figure 4.30. With *Random forest* classifier *Oscar Award*, *Golden Globe Award* and *Venice Award* achieved F-measure of 0.668. *Golden Globe Award* achieved the highest F-measure with *Naïve Bayes*. However, *Oscar Award* achieved 0.679 F-measure and *Venice Award* has lower F-measure of 0.665. *Oscar Award*, *Golden Globe Award* and *Venice Award* obtained the same result by scoring F-measure of 0.668 with *Support Vector Machine*. With *Decision Tree,* all these awards achieved the F-measure of 0.668.

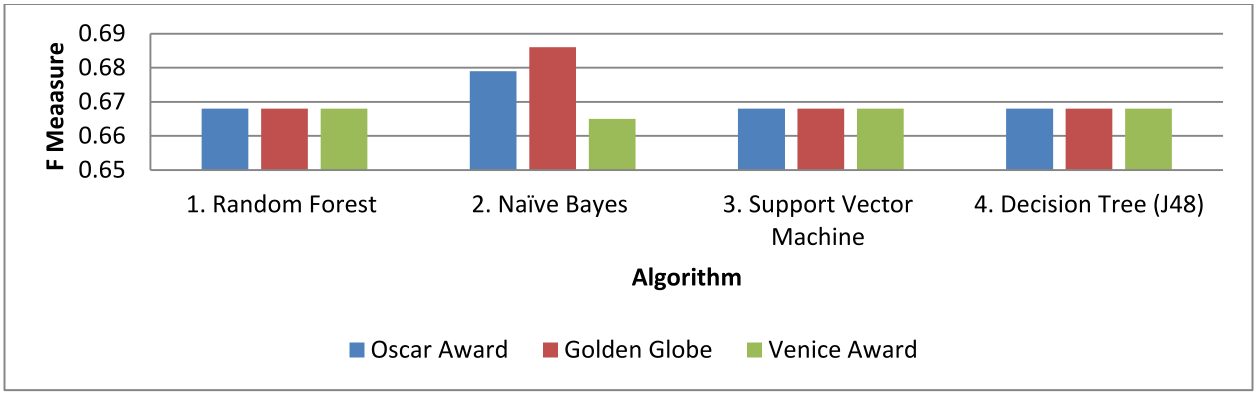


Figure 4.30 Actor 2 - One Feature Analysis Successful and Un-successful Classes

**Findings:** With *Naïve Bayes Classifier*, F-measure of 0.686 has obtained by *Golden Globe Award.*

**(D) One Feature Analysis for Supporting Actor 3**

The results of one feature analysis for supporting actor 3 have been shown in this section.

Table 4.16: One Feature Analysis Supporting Actor 3 Award

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar Award | 0.659 | 0.751 | 0.678 | 0.115 | 0.329 | 0.171 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.114 | 0.329 | 0.17 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **2. Naïve Bayes** | Oscar Award | 0.659 | 0.751 | 0.678 | 0.115 | 0.329 | 0.171 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.095 | 0.173 | 0.067 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **3. Support Vector Machine** | Oscar Award | 0.659 | 0.751 | 0.678 | 0.112 | 0.335 | 0.168 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Oscar Award | 0.591 | 0.769 | 0.668 | 0.115 | 0.329 | 0.171 |
| | Golden Globe | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |

Figure 4.31 shows the results of all classes. With *Random forest* classifier *Oscar Award* achieved highest F-measure of 0.171. *Golden Globe Award* achieved F-measure of 0.17

and *Venice Award* achieved the lowest F-measure of 0.168. Oscar Award achieved highest F-measure with *Naïve Bayes* i.e. 0.171. However, *Golden Globe Award* achieved the lowest F-measure of 0.067 and *Venice Award* has F-measure of 0.168. *Oscar Award, Golden Globe Award* and *Venice Award* obtained the same result by scoring F-measure of 0.168 with *Support Vector Machine*. With *Decision Tree, Oscar Awards* obtained highest F-measure of 0.171 and *Golden Globe Award* and *Venice Award* achieved the F-measure of 0.168.

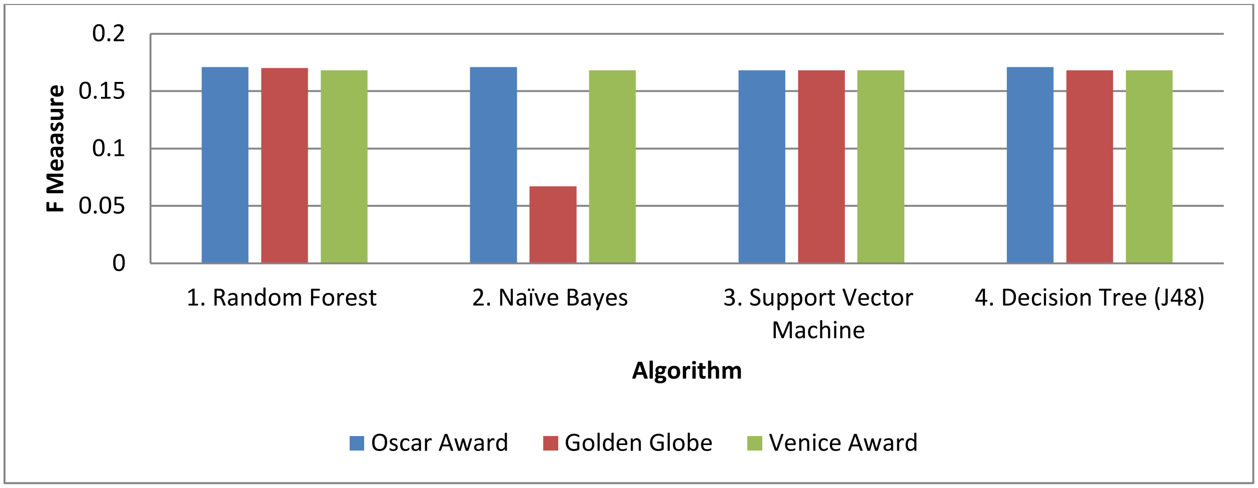


Figure 4.31 Actor 3 - One Feature Analysis Class A to Class I

**Findings:** It has been evaluated that *Oscar Award* has highest F-measure of 0.171 with *Random Forest, Naïve Bayes and Decision Tree.*

In Figure 4.32 two classes' results are shown. With *Random forest* classifier *Oscar Award* achieved the highest F-measure of 0.678 while *Golden Globe Award* and *Venice Award* achieved F-measure of 0.668. *Oscar Award* achieved the highest F-measure with *Naïve Bayes i.e*. 0.678. However, *Golden Globe Award* and *Venice Award* have achieved F-measure of 0.668. *Oscar Award* obtained highest F-measure of 0.678 with *Support Vector Machine*. *Golden Globe Award* and *Venice Award* obtained the same result by scoring F-measure of 0.668. With *Decision Tree*, all these awards achieved the F-measure of 0.668.

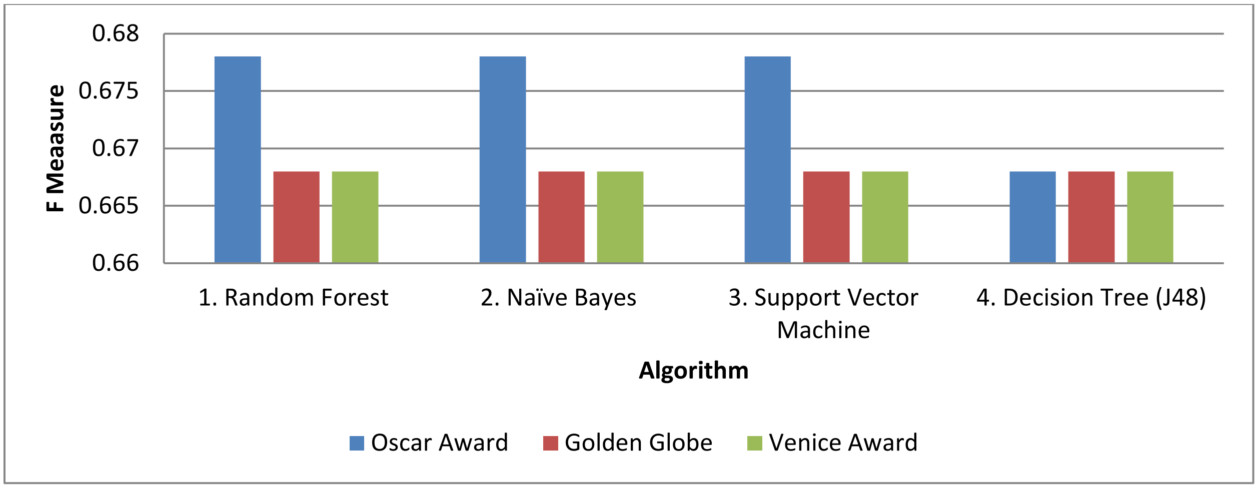


Figure 4.32 Actor 3 - One Feature Analysis Successful and Un-successful Classes

**Findings:** *Oscar Award* obtained the best result by scoring F-measure of 0.678 with *Random Forest*, *Naïve Bayes* and *Support Vector Machine*.

### 4.5.2 Two Feature Analyses

This section evaluates two combined features analysis. More specifically, what remained the F-measure by combining Twitter and Facebook followers, Facebook and Instagram followers and Twitter and Instagram followers?

#### (A) Two Feature Analysis for Director

The results of two feature analysis for director have been shown in this section.

Table 4.17: Two Feature Analysis Director Award

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar+Golden Globe | 0.709 | 0.769 | 0.679 | 0.286 | 0.312 | 0.192 |
| | Oscar+Venice | 0.709 | 0.769 | 0.679 | 0.214 | 0.312 | 0.189 |
| | Golden Globe +Venice Award | 0.751 | 0.775 | 0.691 | 0.201 | 0.347 | 0.194 |
| **2. Naïve Bayes** | Oscar+Golden Globe | 0.628 | 0.746 | 0.666 | 0.163 | 0.312 | 0.191 |
| | Oscar+Venice | 0.649 | 0.757 | 0.672 | 0.163 | 0.318 | 0.194 |
| | Golden Globe +Venice Award | 0.589 | 0.757 | 0.663 | 0.23 | 0.347 | 0.208 |
| **3. Support Vector Machine** | Oscar+Golden Globe | 0.591 | 0.769 | 0.668 | 0.28 | 0.168 | 0.048 |
| | Oscar+Venice | 0.591 | 0.769 | 0.668 | 0.197 | 0.341 | 0.189 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.198 | 0.341 | 0.181 |
| **4. Decision Tree (J48)** | Oscar+Golden Globe | 0.591 | 0.769 | 0.668 | 0.253 | 0.312 | 0.189 |
| | Oscar+Venice | 0.591 | 0.769 | 0.668 | 0.172 | 0.306 | 0.178 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.209 | 0.341 | 0.2 |

Results of two classes are shown in Figure 4.33. With *Random forest* classifier is applied on two features it has been seen that *Golden Globe+ Venice Awards* achieved the highest F-measure of 0.691 while *Oscar+ Golden Globe* and *Oscar+ Venice Awards* achieved the F-measure of 0.679. *Oscar+ Venice Award* achieved the highest F-measure of 0.672 with *Naïve Bayes*. However, *Oscar+ Golden Globe Award* achieved the F-measure of 0.666 and *Golden Globe +Venice Award* scored 0.663 F-measure. *Oscar+ Golden Globe Awards, Oscar+ Venice Awards* and *Golden Globe+ Venice Awards* obtained the same result by scoring F-measure of 0.668. With *Decision Tree,* again these two features achieved the same F-measure of 0.668.

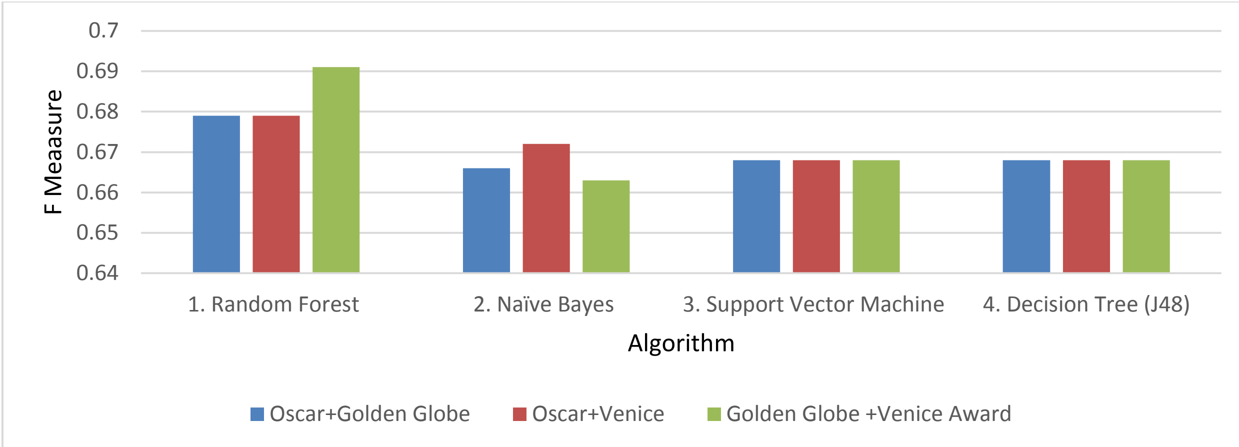


Figure 4.33 Director Two Feature Analysis Successful and Un-successful Classes

**Findings:** With *Random Forest Classifier, Golden Globe+ Venice Awards* obtained the best result by scoring F-measure of 0.691

Results of analyzing the two features with all classes are shown in Figure 4.34. When *Random forest* classifier is applied on two features, it has been seen that *Golden Globe+ Venice Awards* achieved the highest F-measure of 0.194 while *Oscar+ Golden Globe* achieved 0.192 F-measure and *Oscar+ Venice Awards* achieved the lowest F-measure of 0.192. *Golden Globe +Venice Award* achieved the highest F-measure of 0.208 with *Naïve Bayes*. However, *Oscar + Venice Award* achieved the F-measure of 0.194 and *Oscar+ Golden Globe Award* scored 0.191 F-measure. *Oscar + Venice Awards* obtained the highest result by scoring F-measure of 0.189 with *Support Vector Machine*. *Golden Globe +Venice Award* achieved F-measure of 0.181, however; *Oscar+ Golden Globe Awards* scored the lowest F-measure of 0.048. With *Decision Tree, Oscar + Golden Globe Awards* scored the highest F-measure of 0.189. *Oscar + Venice Awards* have scored 0.178 F-measure and *Golden Globe +Venice Award* achieved the lowest F-measure of 0.2.

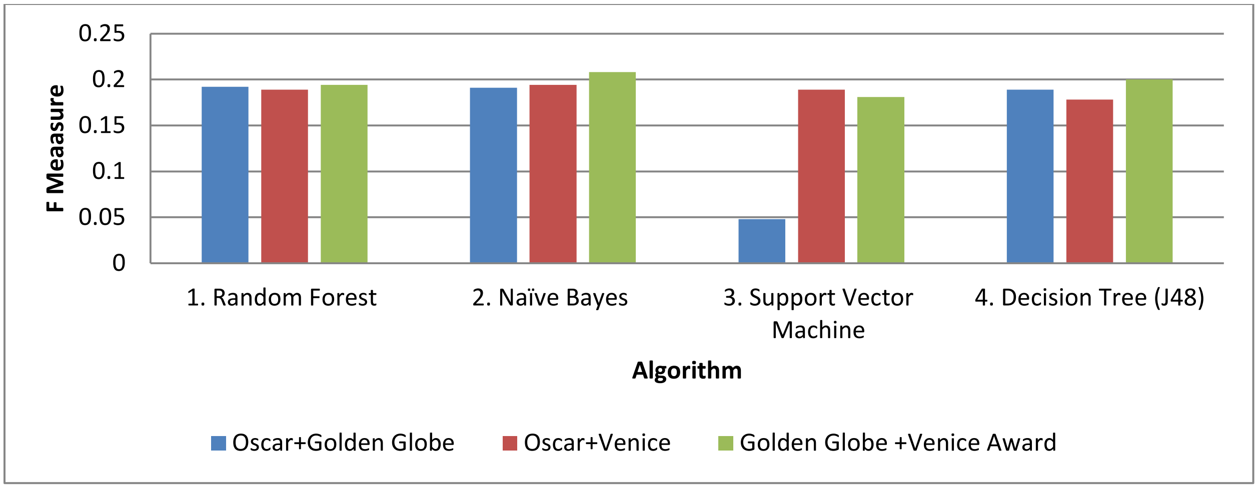


Figure 4.34 Director Two Feature Analysis Class A to Class I

**Findings:** Highest F-measure of 0.208 has been obtained by *Golden Globe+ Venice Awards Naïve Bayes*.

**(B) Two Feature Analysis for Lead Actor 1**

The results of two feature analysis for lead actor 1 have been shown in this section.

Table 4.18: Two Feature Analysis Lead Actor Award

| **Classifier** | | **Two Output Classes** | | | **All Classes** | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar+Golden Globe | 0.591 | 0.769 | 0.668 | 0.2 | 0.341 | 0.183 |
| | Oscar+Venice | 0.591 | 0.769 | 0.668 | 0.112 | 0.318 | 0.165 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.2 | 0.347 | 0.193 |
| **2. Naïve Bayes** | Oscar+Golden Globe | 0.751 | 0.775 | 0.691 | 0.2 | 0.341 | 0.183 |
| | Oscar+Venice | 0.669 | 0.763 | 0.676 | 0.112 | 0.329 | 0.168 |
| | Golden Globe +Venice Award | 0.687 | 0.763 | 0.685 | 0.199 | 0.335 | 0.181 |
| **3. Support Vector Machine** | Oscar+Golden Globe | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Oscar+Venice | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Oscar+Golden Globe | 0.591 | 0.769 | 0.668 | 0.198 | 0.341 | 0.181 |
| | Oscar+Venice | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |

Figure 4.35 show the results of two classes. With *Random forest* classifier *Oscar+ Golden Globe*, *Oscar+ Venice Awards* and *Golden Globe+ Venice Awards* achieved the same F-measure of 0.668. *Oscar+ Golden Globe Award* achieved the highest F-measure of 0.691 with *Naïve Bayes*. However, *Golden Globe+ Venice Award* achieved the F-measure of 0.685 and *Oscar +Venice Award* scored lowest F-measure of 0.676. *Oscar+ Golden Globe Awards*, *Oscar+ Venice Awards* and *Golden Globe+ Venice Awards* obtained the same result by scoring F-measure of 0.668. With *Decision Tree,* again these two features achieved the same F-measure of 0.668.

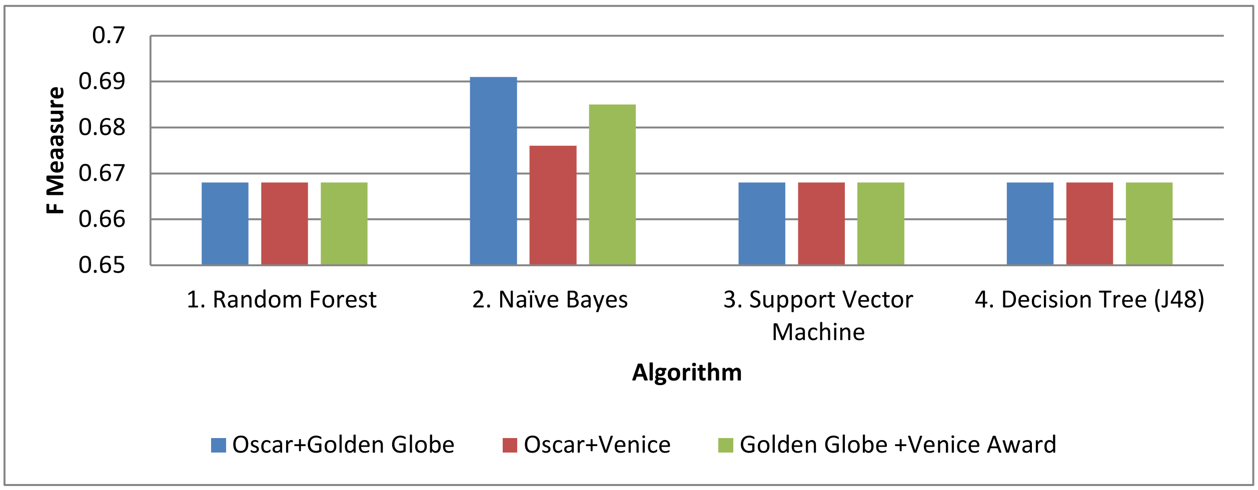


Figure 4.35 Actor one- Two Feature Analysis Successful and Un-successful Classes

**Findings:** *Oscar+ Golden Globe Awards* obtained the best result with *Naïve Bayes Classifier* i.e. 0.691.

Results of two features with all classes are shown in Figure 4.36. With *Random forest* classifier *Golden Globe+ Venice Awards* achieved the highest F-measure of 0.193 while *Oscar+ Golden Globe* achieved 0.183 F-measure and *Oscar+ Venice Awards* achieved the lowest F-measure of 0.165. *Oscar+ Golden Globe Award* achieved the highest F-measure of 0.183 with *Naïve Bayes*. However, *Golden Globe+ Venice Award* achieved the F-measure of 0.181 and *Oscar+ Venice Award* scored lowest F-measure of 0.168. *Oscar+ Golden Globe Awards*, *Oscar+ Venice Awards* and *Golden Globe +Venice Award* obtained the same result by scoring F-measure of 0.668 with *Support Vector*

*Machine*. With *Decision Tree, Oscar+ Golden Globe Awards* scored the highest F-measure of 0.181. *Oscar+ Venice Awards* and *Golden Globe +Venice Award* achieved the same F-measure of 0.668.

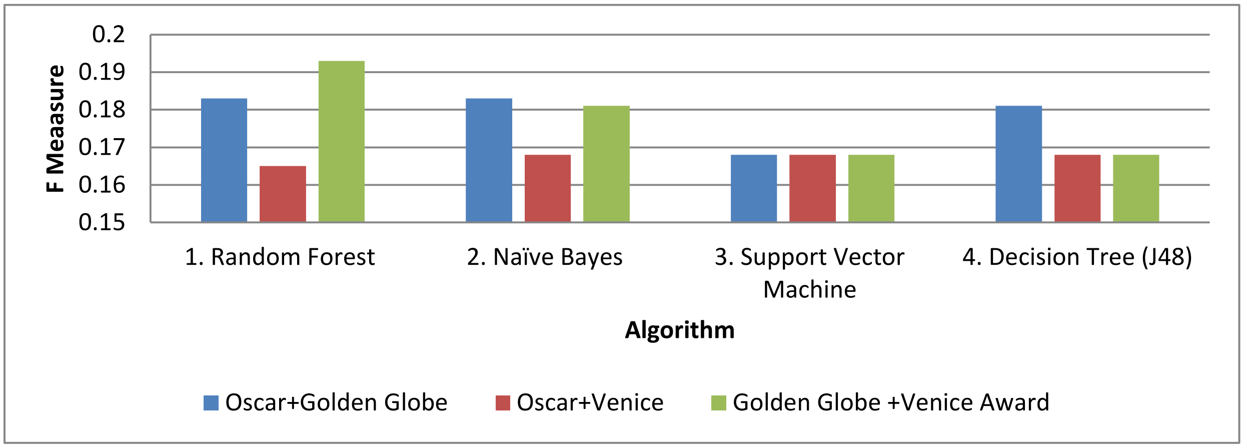


Figure 4.36 Actor one- Two Feature Analysis Class A to Class I

**Findings:** F-Measure of 0.193 has been obtained with *Random Forest* by *Golden Globe+ Venice Awards*.

**(C) Two Feature Analysis for Supporting Actor 2**

The results of two feature analysis for supporting actor 2 have been shown in this section.

Table 4.19: Two Feature Analysis Supporting Actor 2 Award

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar+Golden Globe | 0.591 | 0.761 | 0.668 | 0.201 | 0.324 | 0.182 |
| | Oscar+Venice | 0.591 | 0.761 | 0.668 | 0.25 | 0.335 | 0.194 |
| | Golden Globe +Venice Award | 0.591 | 0.761 | 0.668 | 0.213 | 0.329 | 0.187 |
| **2. Naïve Bayes** | Oscar+Golden Globe | 0.69 | 0.751 | 0.7 | 0.15 | 0.329 | 0.192 |
| | Oscar+Venice | 0.65 | 0.734 | 0.675 | 0.154 | 0.335 | 0.189 |
| | Golden Globe +Venice Award | 0.664 | 0.746 | 0.682 | 0.137 | 0.324 | 0.176 |
| **3. Support Vector Machine** | Oscar+Golden Globe | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Oscar+Venice | 0.591 | 0.769 | 0.668 | 0.198 | 0.341 | 0.181 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Oscar+Golden Globe | 0.591 | 0.769 | 0.668 | 0.201 | 0.335 | 0.184 |
| | Oscar+Venice | 0.591 | 0.769 | 0.668 | 0.25 | 0.335 | 0.194 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |

The results of two classes (Successful and Un-successful) are shown in the Figure 4.37. With *Random forest* classifier *Oscar + Golden Globe*, *Oscar + Venice Awards* and *Golden Globe + Venice Awards* achieved the same F-measure of 0.668. *Golden Globe+ Venice Award* achieved the highest F-measure of 0.682 with *Naïve Bayes*. However, *Oscar+ Venice Award* achieved the F-measure of 0.675 and *Oscar + Golden Globe Award* scored lowest F-measure of 0.7. *Oscar + Golden Globe Awards*, *Oscar + Venice Awards* and *Golden Globe+ Venice Awards* obtained the same result by scoring F-measure of 0.668 with *Support Vector Machine*. With *Decision Tree,* again these two features achieved the same F-measure of 0.668.

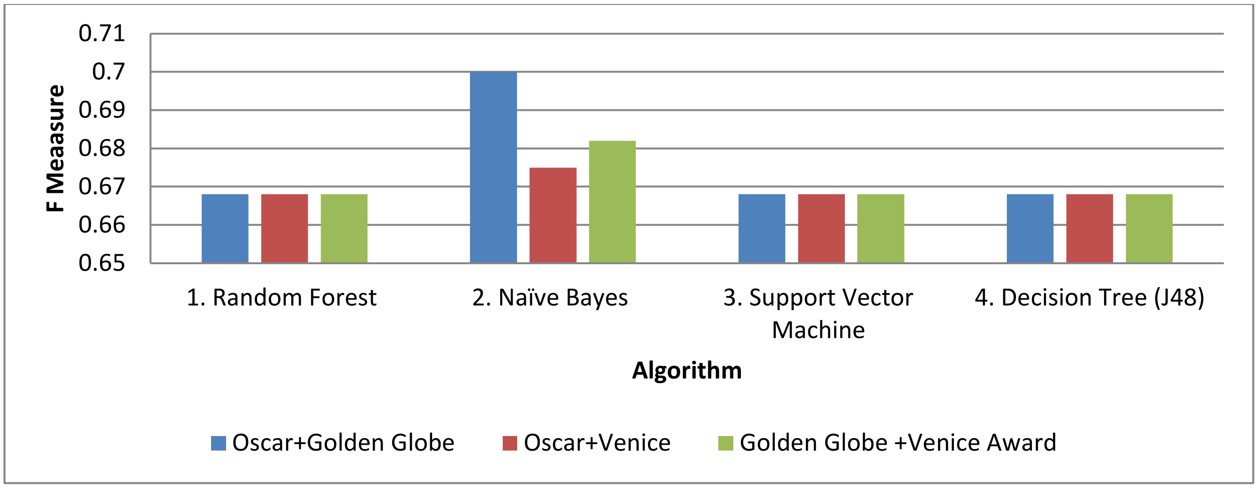


Figure 4.37 Actor 2- Two Features Analysis Successful and Un-successful Classes

**Findings:** With *Naïve Bayes*, *Oscar+ Golden Globe Awards* obtained the highest F-measure of 0.7.

Figure 4.38 demonstrate the results of two classes. With *Random forest* classifier *Oscar+ Venice Awards* achieved the highest F-measure of 0.194. *Golden Globe+ Venice Award* achieved 0.187 and *Oscar + Golden Globe Award* scored 0.182. *Oscar + Golden Globe Award* achieved the highest F-measure of 0.192 with *Naïve Bayes*. However, *Oscar + Venice Award* achieved the F-measure of 0.189 and *Golden Globe+ Venice Award* scored lowest F-measure of 0.176. *Oscar+ Venice Awards* have scored F-measure of 0.181 with *Support Vector Machine*. However, *Oscar+ Golden Globe Awards* and *Golden Globe+ Venice Awards* obtained the same result by scoring F-measure of 0.168 with *Support Vector Machine*. With *Decision Tree, Oscar+ Venice Awards* have scored highest F-measure of 0.194. *Oscar+ Golden Globe Award* achieved the F-measure of

0.184, however; *Golden Globe + Venice Awards* achieved the lowest F-measure of 0.168.

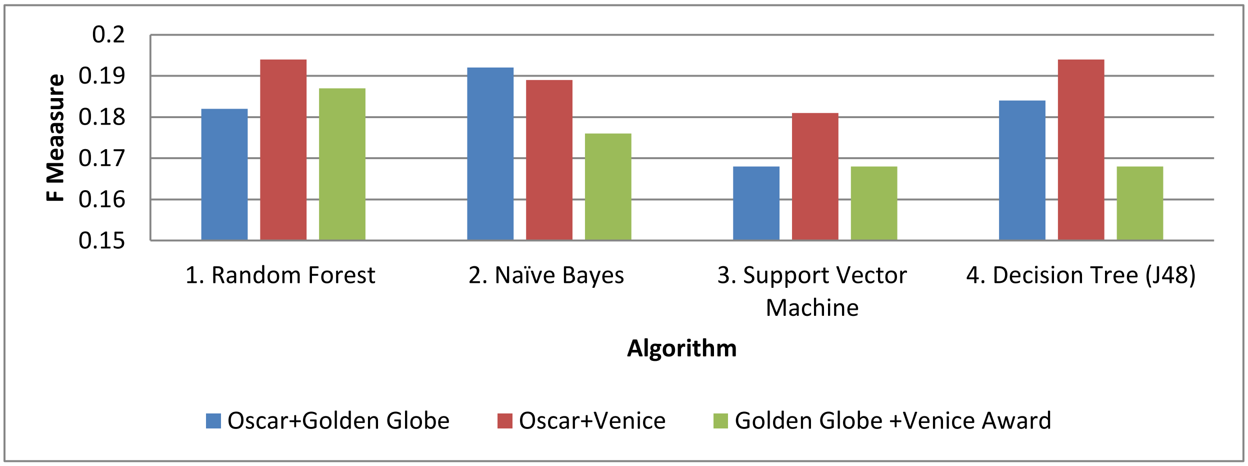


Figure 4.38 Actor 2- Two Feature Analysis Class A to Class I

**Findings:** *Oscar + Venice Awards* achieved the best result by scoring F-measure of 0.194 with *Random Forest* and *Decision Tree*.

### (D) Two Feature Analysis for Supporting Actor 3

The results of two feature analysis for supporting actor 3 have been shown in this section.

Table 4.20: Two Feature Analysis Supporting Actor 3 Award

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar+Golden Globe | 0.826 | 0.775 | 0.682 | 0.134 | 0.329 | 0.179 |
| | Oscar+Venice | 0.659 | 0.751 | 0.678 | 0.115 | 0.329 | 0.171 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.123 | 0.335 | 0.177 |
| **2. Naïve Bayes** | Oscar+Golden Globe | 0.659 | 0.751 | 0.678 | 0.101 | 0.168 | 0.065 |
| | Oscar+Venice | 0.659 | 0.751 | 0.678 | 0.115 | 0.329 | 0.171 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.095 | 0.173 | 0.067 |
| **3. Support Vector Machine** | Oscar+Golden Globe | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Oscar+Venice | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Oscar+Golden Globe | 0.826 | 0.775 | 0.682 | 0.114 | 0.329 | 0.169 |
| | Oscar+Venice | 0.591 | 0.769 | 0.668 | 0.115 | 0.329 | 0.171 |
| | Golden Globe +Venice Award | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |

Figure 4.39 show the analysis of two classes. With *Random forest* classifier *Oscar+ Golden Globe* achieved the highest F-measure of 0.682. *Oscar+ Venice Awards* have scored 0.678 and *Golden Globe+ Venice Awards* achieved the F-measure of 0.668. *Oscar +Golden Globe Award* and *Oscar+ Venice Award* achieved the highest F-measure of 0.672 with *Naïve Bayes*. However, *Golden Globe+ Venice Award* scored lowest F-measure of 0.668. *Oscar+ Golden Globe Awards*, *Oscar+ Venice Awards* and *Golden Globe+ Venice Awards* obtained the same result by scoring F-measure of 0.668 with *Support Vector Machine*. With *Decision Tree*, *Oscar+ Golden Globe Awards* achieved high F-measure of 0.682 while *Oscar+ Venice Award* and *Golden Globe+ Venice Awards* achieved the same F-measure of 0.668.

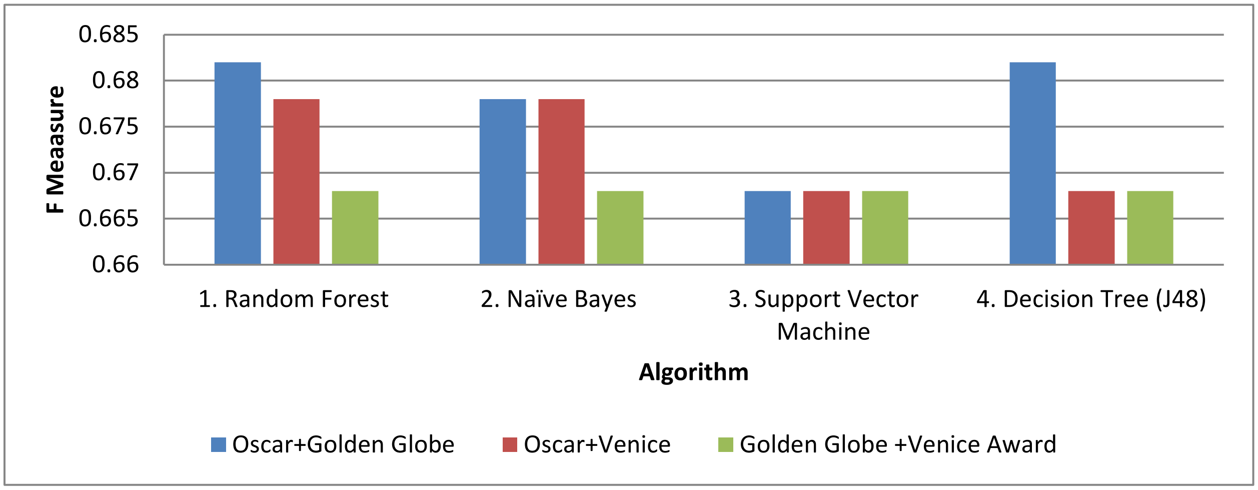


Figure 4.39 Actor 3- Two Feature Analysis Successful and Un-successful Classes

**Findings:** *Oscar+ Golden Globe Awards* obtained the highest F-measure with *Random Forest* and *Decision Tree* i.e. 0.682.

Results analysis of all classes are shown in Figure 4.40. With *Random forest* classifier *Oscar+ Golden Globe Awards* achieved the highest F-measure of 0.179. *Golden Globe+ Venice Award* achieved 0.177 and *Oscar+ Venice Award* scored 0.171. *Oscar+ Venice Award* achieved the highest F-measure of 0.171 with *Naïve Bayes*. However, *Golden Globe+ Venice Award* achieved the F-measure of 0.067 and *Oscar+ Golden Globe Award* scored lowest F-measure of 0.065. *Oscar+ Golden Globe Awards*, *Oscar+ Venice*

*Awards* and *Golden Globe+ Venice Awards* obtained the same result by scoring F-measure of 0.168 with *Support Vector Machine*. With *Decision Tree, Oscar+ Venice Awards* have scored highest F-measure of 0.171. *Oscar+ Golden Globe Award* achieved the F-measure of 0.169 however; *Golden Globe+ Venice Awards* achieved the F-measure of 0.168.

**Findings:** With *Random Forest Classifier*, 0.179 F-measure has obtained by *Oscar+ Golden Globe Awards*.

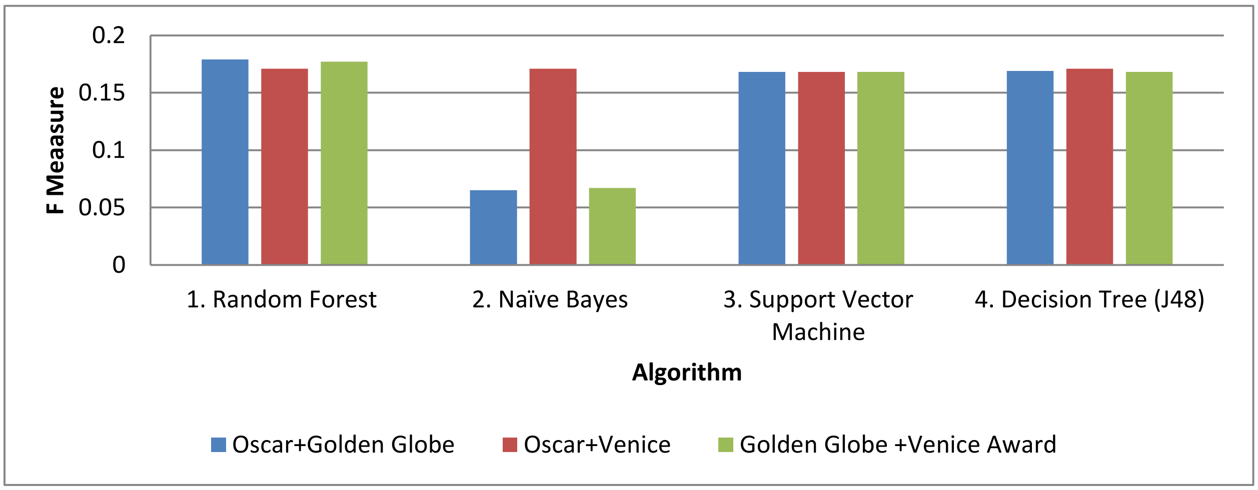


Figure 4.40 Actor 3- Two Feature Analysis Class A to Class I

### 4.5.3 Three Feature Analyses

This section evaluates the results of all three features. To be concrete, we want to identify that when all social media features are combines, then whether we are able to achieve better F-measure or not?

**(A) Three Feature Analysis for Director**

The results of three feature analysis for director have been shown in this section.

Table 4.21: Three Feature Analysis Director Award

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar+Golden Globe+Venice | 0.774 | 0.78 | 0.704 | 0.283 | 0.324 | 0.202 |
| **2. Naïve Bayes** | Oscar+Golden Globe+Venice | 0.649 | 0.757 | 0.672 | 0.162 | 0.318 | 0.196 |
| **3. Support Vector Machine** | Oscar+Golden Globe+Venice | 0.591 | 0.769 | 0.668 | 0.197 | 0.341 | 0.189 |
| **4. Decisi-on Tree (J48)** | Oscar+Golden Globe+Venice | 0.591 | 0.769 | 0.668 | 0.286 | 0.318 | 0.201 |

With *Random forest* classifier, it has been seen that *Oscar+ Golden Globe+ Venice Awards* achieved the highest F-measure of 0.704. With *Naïve Bayes*, F-measure of 0.672 is achieved. While *Support Vector Machine* and *Decision Tree* obtained the same F-measure of 0.668 as shown in Figure 4.41.

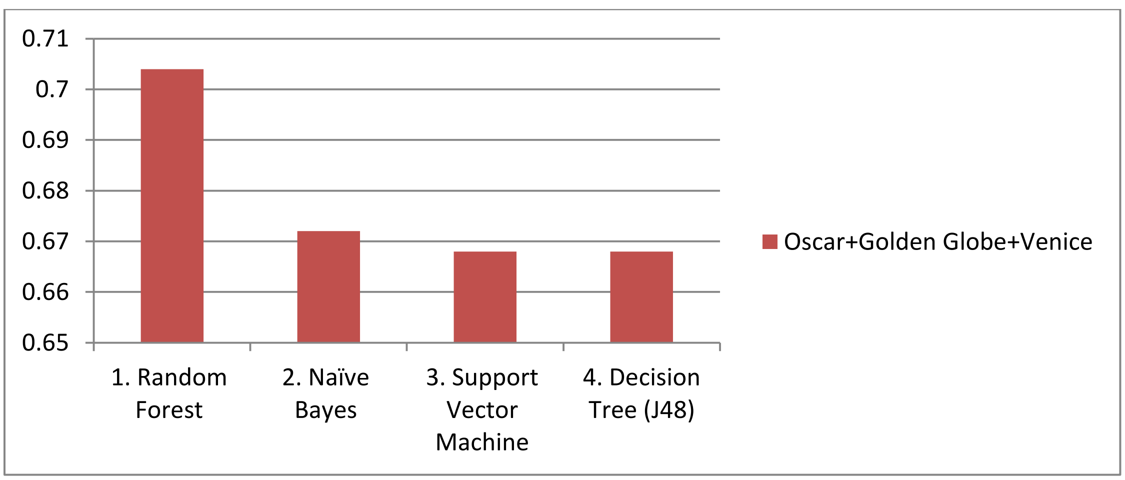


Figure 4.41 Director Three Feature Analysis with Successful and Un-successful Classes

**Findings:** *Random Forest Classifier* achieved highest F-measure of 0.704.

Figure 4.42 show the results of all classes. With *Random forest* classifier it has been seen that *Oscar+ Golden Globe+ Venice Awards* achieved the highest F-measure of 0.202. With *Naïve Bayes*, F-measure of 0.196 is achieved. While *Support Vector Machine* achieved F-measure of 0.189 and *Decision Tree* obtained the same F-measure of 0.201.

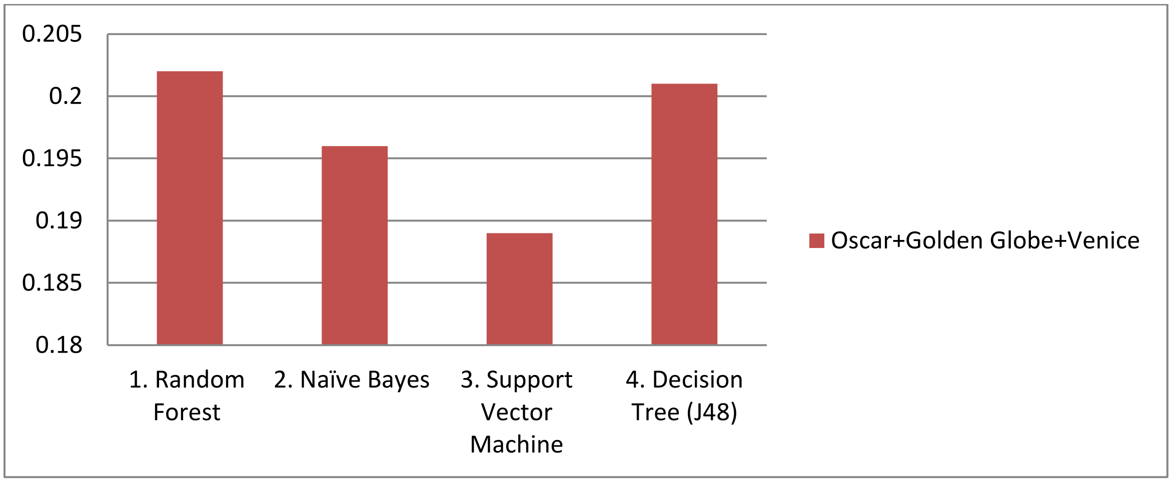


Figure 4.42 Director Three Feature Analysis with Class A to Class I

**Findings:** Highest F-measure of 0.202 obtained by *Random Forest*.

**(B) Three Feature Analysis for Lead Actor 1**

The results of three feature analysis for lead actor 1 have been shown in this section.

Table 4.22: Three Feature Analysis Lead Actor Award

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar+Golden Globe+Venice | 0.709 | 0.769 | 0.679 | 0.218 | 0.335 | 0.2 |
| **2. Naïve Bayes** | Oscar+Golden Globe+Venice | 0.687 | 0.763 | 0.685 | 0.199 | 0.335 | 0.181 |
| **3. Support Vector Machine** | Oscar+Golden Globe+Venice | 0.591 | 0.769 | 0.668 | 0.113 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Oscar+Golden Globe+Venice | 0.591 | 0.769 | 0.668 | 0.198 | 0.341 | 0.181 |

Using two classes, *Random forest* classifier *Oscar+ Golden Globe+ Venice Awards* achieved the F-measure of 0.679. With *Naïve Bayes*, highest F-measure of 0.685 is achieved. While *Support Vector Machine* and *Decision Tree* obtained the same F-measure of 0.668 as shown in Figure 4.43.

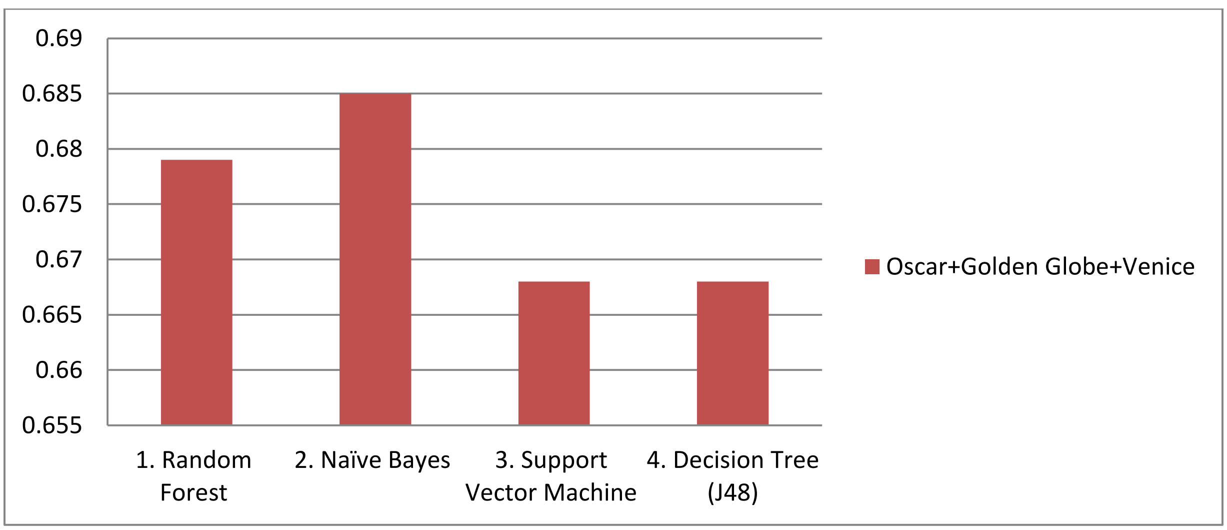


Figure 4.43 Actor one- Three Feature Analysis with Successful and Un-successful Classes

**Findings:** *Naive Bayes Classifier* obtained the best result i.e. 0.685.

Figure 4.44 show the results of all classes. With *Random forest* classifier *Oscar+ Golden Globe+ Venice Awards* achieved the highest F-measure of 0.2. With *Naïve Bayes*, F-measure of 0.181 is achieved. While *Support Vector Machine* achieved F-measure of 0.168 and *Decision Tree* obtained the same F-measure of 0.181.

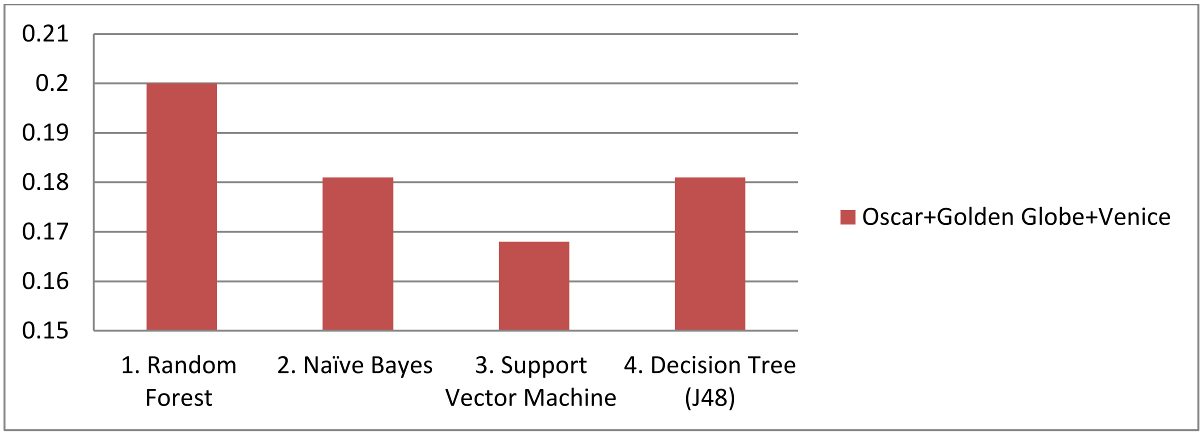


Figure 4.44 Actor one- Three Feature Analysis with Class A to Class I

**Findings:** Best result by scoring F-measure of 0.2 has obtained by *Random Forest.*

### (C) Three Feature Analysis for Supporting Actor 2

The results of three feature analysis for supporting actor 2 have been shown in this section.

Table 4.23: Three Feature Analysis Supporting Actor 2 Award

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar+Golden Globe+Venice | 0.591 | 0.769 | 0.668 | 0.219 | 0.329 | 0.192 |
| **2. Naïve Bayes** | Oscar+Golden Globe+Venice | 0.682 | 0.746 | 0.696 | 0.15 | 0.329 | 0.193 |
| **3. Support Vector Machine** | Oscar+Golden Globe+Venice | 0.591 | 0.769 | 0.668 | 0.198 | 0.341 | 0.181 |
| **4. Decision Tree (J48)** | Oscar+Golden Globe+Venice | 0.591 | 0.769 | 0.668 | 0.251 | 0.341 | 0.196 |

Figure 4.45 show the results of two classes. With *Random forest* classifier *Oscar+ Golden Globe+ Venice Awards* achieved the F-measure of 0.668. With *Naïve Bayes*, highest F-measure of 0.696 is achieved. While *Support Vector Machine* and *Decision Tree* obtained the same F-measure of 0.668.

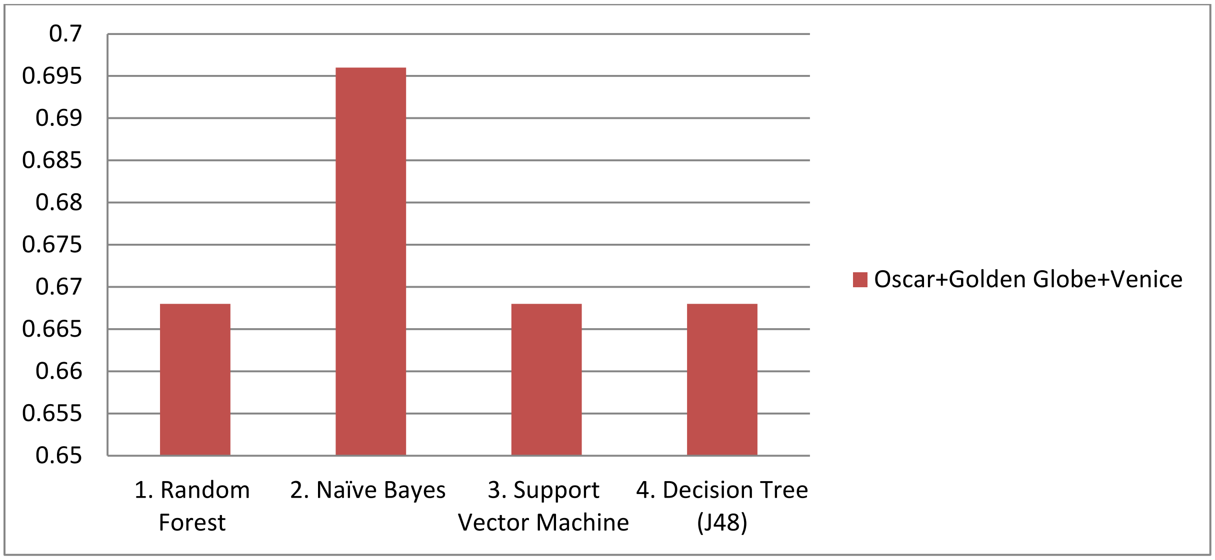


Figure 4.45 Actor 2- Three Feature Analysis with Successful and Un-successful Classes

**Findings:** *Naïve Bayes* obtained the best F-measure of i.e. 0.696.

Figure 4.46 show the results of all classes. With *Random Forest* classifier *Oscar+ Golden Globe+ Venice Awards* achieved the F-measure of 0.192. With *Naïve Bayes*, F-measure of 0.193 is achieved. While *Support Vector Machine* achieved F-measure of

0.181 and *Decision Tree* obtained the highest F-measure of 0.196.

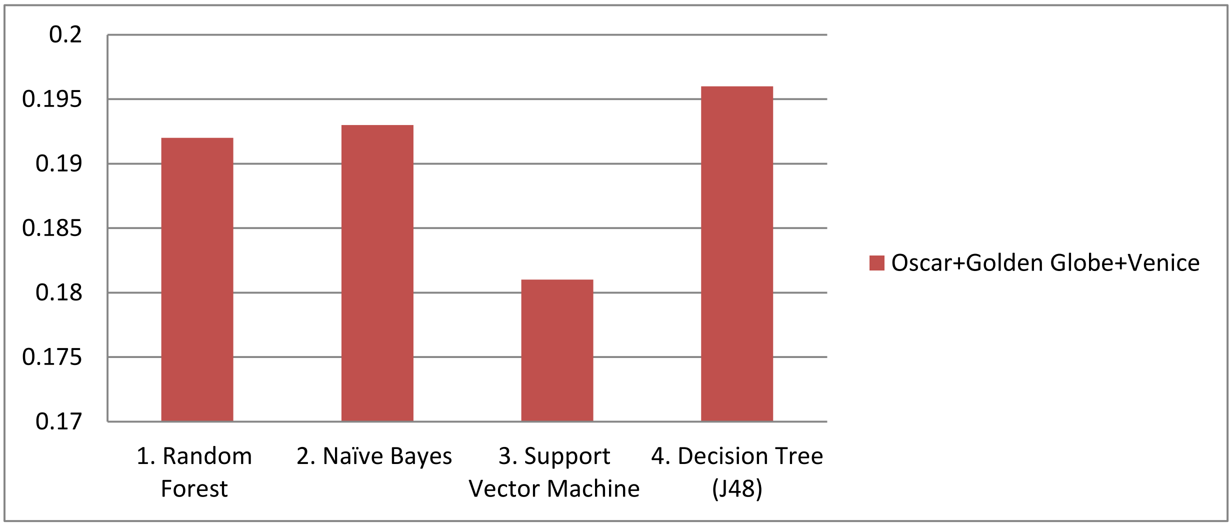


Figure 4.46 Actor 2- Three Feature Analysis with Class A to Class I

**Findings:** *Decision Tree* obtained the best result by scoring F-measure of 0.196.

**(D) Two Feature Analysis for Supporting Actor 3**

The results of three feature analysis for supporting actor 3 have been shown in this section.

Table 4.24: Three Feature Analysis Supporting Actor 3 Award

| Classifier | | Two Output Classes | | | All Classes | | |
|---|---|---|---|---|---|---|---|
| | **Feature Name** | **Precision** | **Recall** | **F-Measure** | **Precision** | **Recall** | **F-Measure** |
| **1. Random Forest** | Oscar+Golden Globe+Venice | 0.826 | 0.775 | 0.682 | 0.156 | 0.335 | 0.179 |
| **2. Naïve Bayes** | Oscar+Golden Globe+Venice | 0.659 | 0.751 | 0.678 | 0.101 | 0.168 | 0.065 |
| **3. Support Vector Machine** | Oscar+Golden Globe+Venice | 0.591 | 0.769 | 0.668 | 0.112 | 0.335 | 0.168 |
| **4. Decision Tree (J48)** | Oscar+Golden Globe+Venice | 0.826 | 0.775 | 0.682 | 0.156 | 0.335 | 0.179 |

Figure 4.47 show the results of two classes. With *Random forest* classifier *Oscar+ Golden Globe+ Venice Awards* achieved the F-measure of 0.682. With *Naïve Bayes*, F-measure of 0.678 is achieved. While *Support Vector Machine* the F-measure of 0.668 is achieved and with *Decision Tree* obtained the F-measure of 0.682.

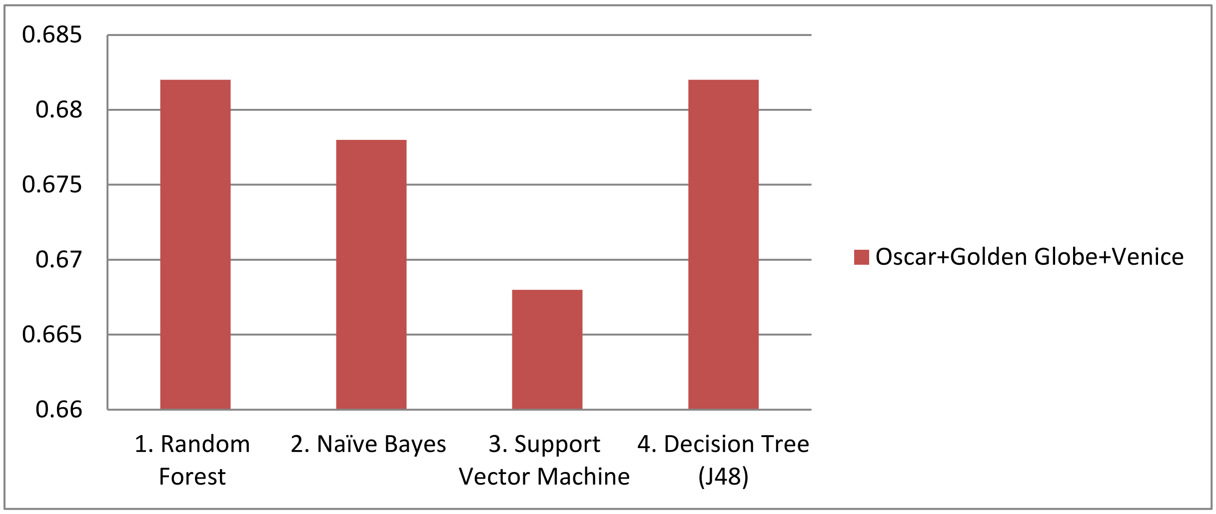


Figure 4.47 Actor 3- Three Feature Analysis Successful and Un-successful classes

**Findings:** F-measure of 0.682 has obtained by *Random Forest Classifier* and *Decision Tree.*

Figure 4.48 demonstrate that using all classes, *Random forest* classifier *Oscar+ Golden Globe+ Venice Awards* achieved the F-measure of 0.179. With *Naïve Bayes*, lowest F-measure of 0.065 is achieved. While *Support Vector Machine* achieved F-measure of

0.168 and *Decision Tree* obtained the highest F-measure of 0.179.

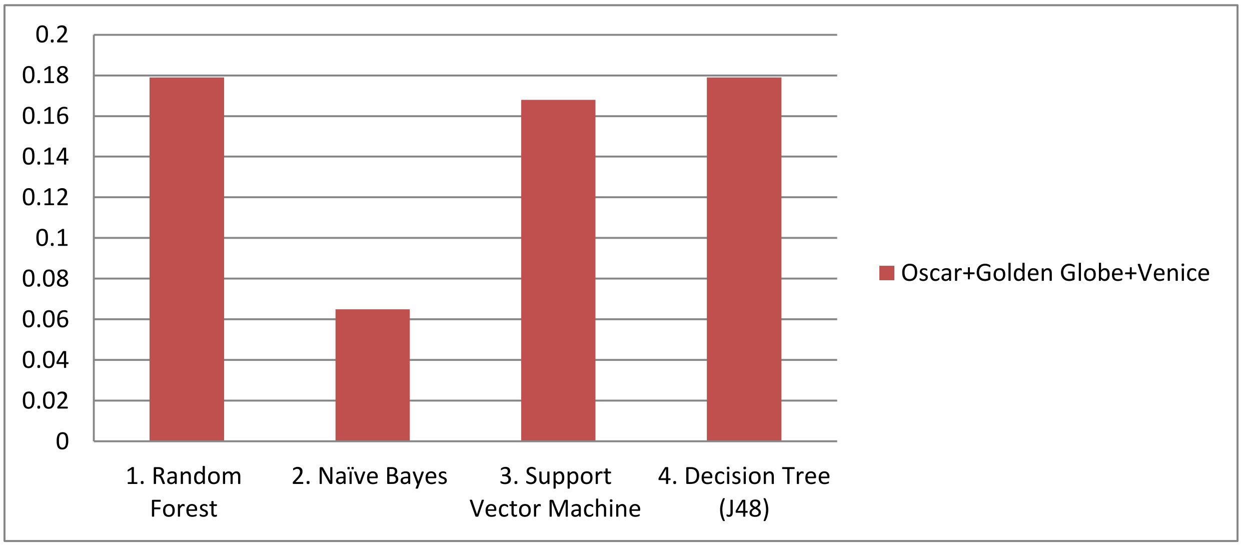


Figure 4.48 Actor 3- Three Feature Analysis with Class A to Class I

**Findings:** *Random Forest* and *Decision Tree* obtained the highest F-measure of 0.179.

| **Feature Analysis** | **All Classes** | | **Two Class** | |
|---|---|---|---|---|
| | **Feature Name** | **F1 Measure** | **Feature Name** | **F1 Measure** |
| **One Feature Analysis** | Golden Globe | 0.204 | Golden Globe | 0.691 |
| **Two Feature Analysis** | Golden Globe+ Venice | 0.208 | Oscar+ Golden Globe | 0.7 |
| **Three Feature Analysis** | Oscar+ Golden Globe+ Venice | 0.202 | Oscar+ Golden Globe+ Venice | 0.704 |

Table 4.25: Results Conclusion

### 4.5.4 Conclusion

This chapter has evaluated all features of social media and awards to comprehensively conclude which feature achieves the best results. There were number of important findings out of these experiments. Such important findings have been discussed below:

1. When all 9 revenue classes are used for classification, the F-Measure drops dramatically ranging from 0.15 to 0.2. This means that all of the evaluated features do not hold a potential to predict the classes on a fine grained level.

2. When two classes are predicted i.e. Successful movies and Un-successful movies, then reasonable F-measure has been achieved ranging from . This means that at least the evaluated features have a possibility to predict the overall category of the movie. Such classification is more important than classifying movies into different revenue ranges. As it does not matter a lot that a movie at hand will go into which range of revenue. However, this is very important for someone to know whether it would be a Successful movie or Unsuccessful movie and knowing such information before even signing the movie, is a great asset for the investor. In this context, the proposed features have scored good F-measure.

3. In one feature analysis, it has been evaluated that Instagram obtained the best results by scoring F-measure of 0.733. Using two feature analyses, Twitter + Instagram achieved the highest F-measure of 0.753. Random Forest classifier has scored 0.781 F-measure using three feature analyses. It means that Instagram and Random Forest have high potential to predict the success with highest F-measure.

4. Similarly, if we look at the awards section it has been clear that using one feature analysis, Golden Globe award has scored highest F-measure of 0.696 and using two feature analyses, Oscar + Golden Globe predictive power increases with 0.70. Three feature analysis results show that Random Forest got highest F-measure among all i.e. 0.704.

# Chapter 5. Conclusion and Future Work

## 5.1 Conclusion

The film industry is considered to be one of huge industries, having huge budget spending and high revenue risks. Among all the worlds film industry, Hollywood produces a large number of movies every year which involves billions of dollars. It is considered to be a high risk endeavor as it faces billions of revenue loss annually. The majority of the movie failed to even recover their production budget. This implies that every movie associated with Hollywood industry carries huge risk and can either earn profit or result in losing billions of dollars for a studio in a year. Knowing this reality, stakeholders associated with the Hollywood industry are prominently interested in an expert system which can forecast the movie revenue at pre-production phase.

Besides the high risk involved in this business, marketing or movie promotions also play an important role in convincing the audience to come and watch the movie. It is seen through different research studies that large numbers of movies are unable to convince the viewers to be watched on a yearly basis, making it impossible for them to even generate their production cost. All these factors collectively are enough to convince the researchers and movie industry stakeholders, the need to have such expert systems which can predict success of a movie in terms of estimated revenue against each planned movie.

After analyzing forty state of art paper, we found that most of them targeted post-production phase or have low prediction F-measure. The variable used in those researches is time dependent and is only getable when story, director and cast are finalized. The invested money already been spent at the time these models give a prediction. These models have a limited scope and non-ability to reduce revenue loss risk. The use of historical data related to film industry should be used to predict the success of the movie. The historical data include the past performance of the movie cast which can be estimated by seeing the number of awards won by the lead cast member.

Moreover, social media cast popularity can also be taken into account to foresee its impact toward movie revenue.

We address these issues in our research by forecasting the movie success at the pre-production phase using purely historical and time independent data. We have comprehensive, evaluated the analytical power social media (Facebook, Twitter & Instagram) and prestigious awards (Oscar, Golden Globe and Venice Award) won by the leading cast of the Hollywood movie. The Twitter is used by a different researcher, however; less significant work has been done using combination social media powerhouses (Facebook, Twitter and Instagram). The key point of this research is to analyze movie lead cast social media (Facebook, Twitter & Instagram) popularity and utilize them to predict success of the movie. Moreover, we also explore the predictive power of prestigious awards (Oscar, Golden Globe and Venice Award) won by directors and the leading cast of the Hollywood movie.

In our proposed methodology we have made use of twenty six different features, using a forecasting model to estimate the revenue of a movie. For this purpose the dataset is gathered from different freely available websites (IMDb, Facebook, Instagram and Twitter) and is comprised of over last 10 years (2005-2015). After pre-processing, we classify revenue ranges into two categories. The first categories have nine different ranges of output classes, i.e. from class A to class I and the second category has only two output classes Successful class and Un-successful class. The social media (Facebook, Twitter & Instagram) and prestigious award (Oscar, Golden Globe and Venice Award) have been evaluated as independent features and in combinations as well. The research focus to find which social media platform such as: Facebook, Twitter and Instagram can predict the success of a movie in a better way and to find which prestigious award (Oscar, Golden Globe and Venice Award) have predictive power to forecast the success of a Hollywood movie.

In social media evaluation, when one feature analysis is performed, the best feature among Facebook, Twitter and Instagram is turning out to be Instagram as it achieved the highest F-measure of 0.733 with Successful and Un-successful classes. During two feature analysis, Twitter+ Instagram is considered as the best combination as it achieved the highest F-measure of 0.753. During three feature analysis, it is found that classification model Random Forest obtained the best result by scoring F-measure of 0.781 with Successful and Un-successful classes.

In awards evaluation, Golden Globe Awards achieved the high F-measure of 0.69 with two classes. When the two feature analysis is performed, The Oscar+Golden Globe combination achieved highest F-measure i.e. 0.70. When three features, analysis Random Forest achieved the highest score of 0.704.

## 5.2 Future Work

We set a number of goals for future research as only we tried page likes of Facebook and followers of Twitter & Instagram. More awards can also be added and evaluated.

1. We have chosen four renowned classification machine learning model and each of them has shown different levels of F-measure. It would be an interesting idea to use others models to get better results.

2. Sentiment analysis can also be performed in combination of these twenty six features which can add improvement in the predictive power.

3. Building a hybrid model for making such prediction remain another future goal for the research.

# References


Ainslie, A., Drèze, X. and Zufryden, F., 2005. Modeling movie life cycles and market share. *Marketing Science*, *24*(3), pp.508-517.

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

Baek, H., Ahn, J. and Oh, S., 2014. Impact of tweets on box office revenue: focusing on when tweets are written. *ETRI Journal*, *36*(4), pp.581-590.

Basuroy, S., Chatterjee, S. and Ravid, S.A., 2003a. How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of marketing*, *67*(4), pp.103-117.

Basuroy, S., Chatterjee, S. and Ravid, S.A., 2003b. How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of marketing*, *67*(4), pp.103-117.

Basuroy, S., Desai, K.K. and Talukdar, D., 2006. An empirical investigation of signaling in the motion picture industry. *Journal of Marketing Research*, *43*(2), pp.287-295.

Chamasemani, F.F. and Singh, Y.P., 2011, September. Multi-class support vector machine (SVM) classifiers--an application in hypothyroid detection and classification. In *Bio-Inspired Computing: Theories and Applications (BIC-TA), 2011 Sixth International Conference* pp. 351-356. IEEE.

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

Delen, D. and Sharda, R., 2010. Predicting the financial success of hollywood movies using an information fusion approach. *Indus Eng J*, *21*(1), pp.30-37.

Dellarocas, C., Awad, N. and Zhang, M., 2005. Using online ratings as a proxy of word-of-mouth in motion picture revenue forecasting. *Smith School of Business, University of Maryland*. , pp.1–32.

Duan, K. and Keerthi, S.S., 2005. Which Is the Best Multiclass SVM Method? An Empirical Study. *Multiple classifier systems*, *3541*, pp.278-285.

El Assady, M., Hafner, D., Hund, M., Jäger, A., Jentner, W., Rohrdantz, C., Fischer, F., Simon, S., Schreck, T. and Keim, D.A., 2013. Visual analytics for the prediction of movie rating and box office performance. *IEEE VAST Challenge USB Proceedings*., pp.3–4.

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), pp.329-354.

Eliashberg, J., Jonker, J.J., Sawhney, M.S. and Wierenga, B., 2000. MOVIEMOD: An implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Science*, *19*(3), pp.226-243.

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), pp.881-893.

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), pp.3176-3193.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B., 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), pp.18-28.

Jain, V., 2013. Prediction of movie success using sentiment analysis of tweets. *The International Journal of Soft Computing and Software Engineering*, *3*(3), pp.308-313.

Eliashberg, J., Elberse, A. and Leenders, M.A., 2006. The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing science*, *25*(6), pp.638-661.

Krauss, J., Nann, S., Simon, D., Gloor, P.A. and Fischbach, K., 2008, June. Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis. In *ECIS*, pp. 2026-2037.

Lee, C. and Jung, M., 2014. Predicting Movie Incomes Using Search Engine Query Data. In *International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, pp.45–49.

Litman, B.R., 1983. Predicting success of theatrical movies: An empirical study. *The Journal of Popular Culture*, *16*(4), pp.159-175.

Lloyd, L., Kechagias, D. and Skiena, S., 2005, November. Lydia: A system for large-scale news analysis. In *International Symposium on String Processing and Information Retrieval*, pp. 161-166.

Magerman, D.M., 1995, June. Statistical decision-tree models for parsing. In *Proceedings* *of the 33rd annual meeting on Association for Computational Linguistics*, pp. 276-283.

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), p.e71226.

Mishne, G. and Glance, N.S., 2006, March. Predicting Movie Sales from Blogger Sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 155-158.

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), pp.3176-3193.

Moon, S., Bergey, P.K. and Iacobucci, D., 2010. Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *Journal of Marketing*, *74*(1), pp.108-121.

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

Parimi, R. and Caragea, D., 2013, July. Pre-release box-office success prediction for motion pictures. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 571-585.

Pazzani, M. and Billsus, D., 1997. Learning and revising user profiles: The identification of interesting web sites. *Machine learning*, *27*(3), pp.313-331.

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

Shruti S., Deb Roy, S,. & Zeng,W.(2014). Influence of social media on performance of movies. Colubia,USA: University of Missouri.

de Silva, B. and Compton, R., 2014. Prediction of foreign box office revenues based on wikipedia page activity. *arXiv preprint arXiv:1405.5924*.

Forest, R., 2003. A Classification and Regression Tool for Compound Classification and QSAR Modeling Svetnik, Vladimir; Liaw, Andy; Tong, Christopher; Culberson, J. Christopher; Sheridan, Robert P.; Feuston, Bradley P. *Journal of Chemical Information and Computer Sciences*, *43*(6), pp.1947-1958.

Terry, N., Butler, M. and De'Armond, D.A., 2011. The determinants of domestic box office performance in the motion picture industry. *Southwestern Economic*

*Review*, *32*, pp.137-148.

Wangt, Y., Phillipst, I.T. and Haralick, R., 2001. Automatic table ground truth generation and a background-analysis-based table structure extraction method. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference*, pp. 528-532.

Wu, J.Y. and Pao, Y., Predicting Sentiment from Rotten Tomatoes Movie Reviews.

Zhang, L., Luo, J. and Yang, S., 2009. Forecasting box office revenue of movies with BP neural network. *Expert Systems with Applications*, *36*(3), pp.6580-6587.

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

Zhang, Z., Chai, J., Li, B., Wang, Y., An, M. and Deng, Z., 2015, December. Movie Box Office Inteval Forecasting Based on CART. In *Computational Intelligence and Design (ISCID), 2015 8th International Symposium*, 2, pp. 87-90.