|| Volume 2 || Issue 6 || December 2017|| ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Evaluating Movie Scripts to Point of Green-Lighting Using SVM and Customized Kernel

Ms. Tejaswini B. Muley

P.G Student, Computer Science & Engineering, C.S.M.S.S., Chh. Shahu College of Engineering, Aurangabad, Maharashtra, India.

muleyteju@gmail.com

Abstract— **Entertainment industry want to increasing performance of box office it is related revenues important for movies, when estimated organization decided budget of movie its dependent upon a scripts. In this paper we propose the prediction of box office performance using movie scripts.**

In this paper proposed extract three levels of textual features 1) Genre and content 2) semantics 3) bag of words, Domain knowledge of screen writing for the scripts processing technique of the input and natural language, These textual features and variables define a distance metric across scripts. There are used as a inputs for kernel-based approach to review box office performance. our proposed method prediction box office performance collection is more accurately i.e 29 percent lower mean squared error (MSE)) compared to benchmark methods.

Keywords: *Entertainment industry, text mining, kernel approach, semantics.*

I INTRODUCTION

As large amount of money is invested in movies, movie industry is largely a guesswork based on experts' experience [1][2]. For Movie studios it is hard to choose thousands of scripts to decide which ones to turn into movies. In this proposed system [2] a new approach to help studios evaluate scripts which will then lead to more profitable prior decisions. Here we combines screen writing domain knowledge, natural language processing techniques and statistical learning methods to forecast a movie's return-on-investment [3] based only on textual information available in movie scripts [4]. It test return-on-investment significantly for deciding which scripts to turn into movies, The movie studios and film makers need to assess the box performance of a movie based only on its script and allocated production budget as most post-production drivers of box office performance e.g., actor, actress, director, MPEG rating are unknown at the point of green-lighting when financial commitments have to be made. Usually movie producers rely on a "comps"-based approach [2] to assess the box office potential of a new script. Specifically, they identify around past movies which are "similar" to the script and we use the box office performance of those movies as benchmarks for the revenue potential.

The System contribution is in three parts first that collects and analyzes actual movie scripts. Second display that the portion approach which indicates beats both reversion and tree-based techniques with regards to evaluating film industry execution and third the assessed "emphasize weights" give a few experiences about which printed highlights require specific consideration while distinguishing helpful "comps" for another content. The next section describes an overview of the script data set & how we extract textual information from script and section 3 describes the kernel-based approach and how can we estimate the obtained feature weights. Finally we compare our method with other benchmark methods and present a hypothetical portfolios selection scenario this proposed method can gives lower mean square error.

II RELATED WORK

We relate an incentive in danger (VaR) to mean-difference examination and look at the monetary ramifications of utilizing a mean-VaR [10] display for portfolio determination. When looking at two mean-fluctuation productive portfolios the higher change portfolio may have less VaR. Therefore a proficient portfolio that universally limits VaR may not exist. A strategy is given for the estimation of the Hessian network in the area of the base, required in factual estimation issues. We depict LDA [5], a generative probabilistic model for accumulations of discrete information, for example, content corpora. In this paper, the model of anticipating deals execution of movies is actualized utilizing supposition data mined from surveys and film industry income. Inactive Semantic Analysis (LSA) [12] is a hypothesis and strategy for extricating and speaking to the relevant utilization importance of words by factual calculations connected to an extensive corpus of content (Landauer and Dumais, 1997). The ampleness of LSA's impression of human learning has been built up in an assortment of ways.

III PROPOSED SYSTEM

A) Extracting Textual Features From Movie Scripts

Data is comprised of movies script which are available online we record the Indian box office revenue and production budget from Internet Movie Database(IMDB) [13] and mojo database [14]. The Genre and Content Variables: The "content" factors measure different parts of the story line of content: e.g., what is the premise of the story line? Is the setting commonplace to generally watchers as appeared in Table 1.1 We consider eight conceivable classes. Having ordered a content into one or more genre(s), readers at that point answer an arrangement of 24 "content" inquiries regarding the storyline for each content (as appeared in Table 1), created by [8] in their investigation of movies story lines. These inquiries are basic "yes or no" inquiries that have been distinguished by screenwriting specialists as vital parts of film contents [6], [7], along these lines giving an instructive arrangement of literary highlights.

We normal readers' (0/1) reactions for each inquiry. We recognize that the above system is essentially subjective as it includes human contribution; there is no achievable option accessible as PCs clearly can't yet understand contents. To investigate the degree of subjectivity, we examine the between rater agreement between the two reader Crosswise over sort and substance questions, the two reader give a similar answer (yes/no) around 83 percent of the time, proposing sensible level of understanding.

TABLE 1:-Summary Description of Genre and Content Variables Extracted from Each Script

Variable	Description
Genre	Genre: A movie may belong to number of the following categories: (DRA) Drama, (ROM) Romance, (THR) Thriller, (HOR) Horror, (SCI) Sci-Fi, (COM) Comedy, (ACT) Action, and (FAM) Family.
CLRPREM	Clear Premise: the story has a Clear Premise.
IMPREM	Important Premise: story Premise that is important to audiences.
FAMSET	Familiar Setting: The setting of the story is familiar to audience.
EAREXP	Early Exposition: Information about character comes very early in story.
COAVOID	Coincidence Avoidance: The story follows logical and casual relationship; Coincidence Avoided
INTCON	Interconnected: each scene Description advances the plot and is closely connected to central conflict.
SURP	Surprise: The story contains element of surprise, but logical within context and within its own rules.
ANTICI	Anticipation: The story keeps readers trying to Anticipate what would happen next.
FLHBACK	Flashback: The story contains Flashback sequences.
CLRMOT	Clear Motivation: the hero of story has clear outer motivation (what he/she wants to achieve by end of the movie).
MULDIM	Multi-dimensional Hero: many dimensions of the hero are explored.
HEROW	Hero Weakness: Hero has an inherent Weakness.
STRNEM	Strong Nemesis: There is Strong Nemesis in the story.
SYMHERO	Sympathetic Hero: The hero attracts your sympathy.
LOGIC	Logical Characters: The actions of main character are logical considering their characteristics.
CHARGROW	Character Growth: Hero changes because of the conflict in the story.
BELIEVE	Believable Ending: the ending is Believable.
SUREND	Surprise Ending: the ending carries surprise or not.

B) Semantic Variables

The second layer of literary data catches the structure by which movies content is composed, and gives a "review" of how the last movies will resemble. As appeared in Fig. 1, a content is sorted out into inside/outside scenes and every scene is involved discoursed talked by the distinctive characters. At the scene level, we acquire a gauge of the aggregate number of scenes in the movies, and how regularly the characters interface in inside or outside space. Most movies specialists concur exchanges influence watchers' pleasure [6].

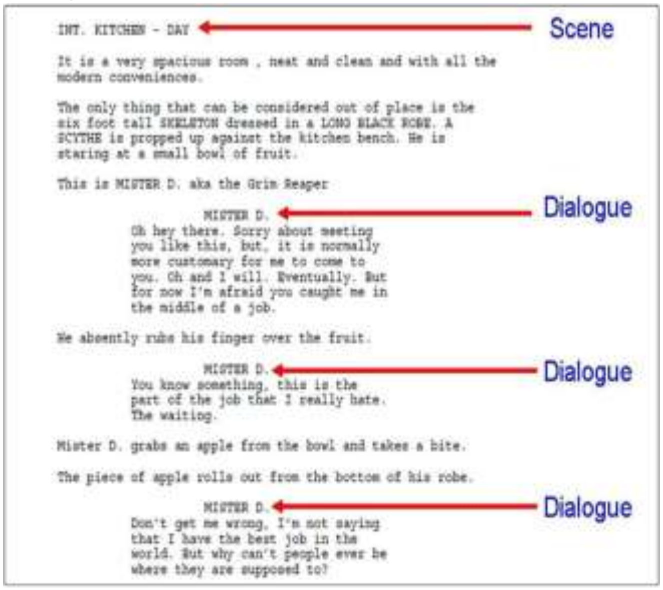
To catch the previously mentioned semantic data, we concentrate on scene factors (i and ii) and exchange factors (iii)- (iv) below

i) NSCENE this variable indicates total number of scenes

ii) INTPREC this variable indicates percentage of interior scenes

iii) NDIAG this variable indicates total number of dialogues.

iv) AVGDIAGLEN this variable indicates average length of dialogues.


Scene
Dialogue
Dialogue
Dialogue

INT. KITCHEN - DAY

It is a very spacious room , neat and clean and with all the modern conveniences.

The only thing that can be considered out of place is the six foot tall SKELETON dressed in a LONG BLACK ROBE. A SCYTHE is propped up against the kitchen bench. He is staring at a small bowl of fruit.

This is MISTER D. aka the Grim Reaper

MISTER D.
Oh hey there. Sorry about meeting you like this, but, it is normally more customary for me to come to you. Oh and I will. Eventually. But for now I'm afraid you caught me in the middle of a job.

He absently rubs his finger over the fruit.

MISTER D.
You know something, this is the part of the job that I really hate. The waiting.

Mister D. grabs an apple from the bowl and takes a bite.

The piece of apple rolls out from the bottom of his robe.

MISTER D.
Don't get me wrong, I'm not saying that I have the best job in the world. But why can't people ever be where they are supposed to?

Figure 1: Semantics structure of a script

C) Bag-of-Words Variables

The third layer of printed data originates from the real words utilized as a part of the content fixed utilizing a "bag-of-words" the representation that is ordinarily utilized as a part of a characteristic language preparing applications. We separate sack of-words data as takes after. In the first place, we dispense with all accentuations, a standard rundown of English names, and "stop words" (e.g., an, an, is, am, are, this, that, him, her) [11]. Next, we utilize a "stemming" calculation [9] to diminish each word to its least difficult frame (e.g., "going" is diminished to go"). We at that point arrange all the extraordinary stemmed words that happen in at least one records to create a word-report lattice. Along these lines following we register a "significance file" for each word, characterized as follows:

$$I_I = (1-\frac{d_i}{D})\mathrm{x}N_I \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where did means the quantity of contents containing the ith word, D means the aggregate number of contents, and Ni is the add up to recurrence of an event of the ith word over all contents. The significance file characterized here is like TFIDF weights. 2 We keep just the best 100 most "critical" words. 3 Finally we perform inert semantic examination (LSA) to additionally decrease the dimensionality of the word-report grid. In light of particular esteem decay (SVD) [12] LSA permits us to list each content by an arrangement of "scores". Since the particular values demonstrate an "elbow" at the two solitary esteem arrangement, we hold two idle semantic scores for each content marked as LS1 and LS2, and utilize them as literary highlights for additionally examinations. Towards that end, we consider the particular vectors comparing to LS1 and LS2 to see which words stack vigorously onto every vector; the rundown of words are appeared in Appendix I, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.organization/10.1109/TKDE.2014.2306681.We hypothesize that LS1 is identified with a specific setting (with words, for example, "robot sheriff", "pilot", "stage", "sword", "boss", "underhanded", "vampire", "tank"), while LS2 gives off an impression of being identified with the style of dialects in the exchanges (e.g., "fxxx", "definitely", "poop", "fella", "going to").

D) Summary and Potential Data Limitations

Outline measurements of every factor in our informational index are appeared in Table 2 Every single literary variable and the (log-) creation a financial plan is then institutionalized and utilized as indicators in a piece based approach to guess film industry execution the proposed method should not be used as an initial screening tool Second the scripts that we collected are “shooting scripts”, i.e. scripts that are used in actual production of the movie Finally, we take note of that our specimen size of 35 contents is littler than the specimen measure normally utilized as a part of a machine learning contemplates due to the little specimen measure, In this manner, our approach straightforwardly addresses the instincts of studio administrators, who are as of now alright with the procedure of getting "comparables" before making forecasts about film industry income.

E) The Kernel-Based Approach:

We utilize the accompanying documentations. Contents in the preparation test are ordered by i=1,…N Each content is involved J particular "highlights," (appeared in Table 2) and is indicated as $\vec{x}=\{x_{ij}\}_{j=1......j}$. along side a "reaction" variable(y_i). we characterize the reaction variable y_i for every film by its (changed) return of venture (ROI). In particular:

$$y_I=\log(BOX\ OFFICE_i/BUDGET_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2)$$

Afterward, we utilize a bit based way to deal with anticipate y_i for a new film, at that point, change it to an expectation of film industry income (Prediction Box $Office_i$=$BUDGET_i$ (e^{y_i})We determine (changed) ROI as the reaction variable in the kernel based strategy on the grounds that such detail presents a few measurable points of interest. Second, the meaning of y_i as of now joins he (straight) impact of log-creation spending plan as a "balance" variable . Third, observationally we find that our determination of y_i in Eq. (2) yields better prescient outcomes contrasted with utilizing film industry income as the reaction variable The separation metric between two perceptions is characterized, in light of (weighted) Euclidean remove, as take

$$d(\overline{\overrightarrow{x_\iota}},\overrightarrow{x_l})=\sqrt{\sum_{j=1}^{j} v_j{}^2(x_{ij}-x_{ij})^2} \quad \text{………………..(3)}$$

where $\vec{v}=\{v_j\}_{j=1\ldots\ldots j}$is a vector of "highlight weights". An indicator variable that has a bigger estimation of v_j is considered more essential when characterizing how comparative two perceptions are. Given Eq. (3), the bit based technique makes forecast for another perception as takes after. Record the new perception by i^*, with a vector of highlights $\overrightarrow{x_{i*}}$; the objective is to foresee the reaction variable y_{i*}.To start with, utilizing Eq. (3),we figure all the pair wise separations d ($\overline{\overrightarrow{x_\iota}},\overrightarrow{x_l}$)between the new perception also, every perception in the preparation database. At that point, we characterize the "weight" of a preparation perception in light of its remove from the new observation:

$$W_{li^*}=\frac{\exp\left(-\theta\, d(\overrightarrow{x_{l*}},\overrightarrow{x_{l*}}\right)}{\sum_l \exp\left(-\theta\, d(\overrightarrow{x_{l*}},\overrightarrow{x_{l*}}\right)} \quad \text{……………..(4)}$$

Note that the scaling parameter u characterizes the degree to which the weights are identified with separations, which will be adjusted in tuning parameter. The bigger u is, the all the more seriously the perceptions that are more not at all like the new perception are down-weighted. the reaction variable yi is anticipated by the weighted entirety of the reaction factors of the preparation observation: $\grave{y}_{i^*}=\sum_l w_{li^*}\, y_l$ …………..(5)

Calibration of the Tuning Parameter θ

We align the "tuning parameters" θ and $\vec{v}$ utilizing a mix of area information and cross approval In particular, we isolate our information into two sets: a preparation test of 265 contents covering movies discharged in 2008 or prior, and a holdout test of 35 contents of motion pictures discharged after January 2009. We adjust u following a contention in [12]. Since θ is the "data transfer capacity" parameter of the Gaussian portion in Eq. (4), it bodes well to set θ with the end goal that it is generally in line with the scope of the separations d ($\overline{\overrightarrow{x_\iota}},\overrightarrow{x_l}$). To see this, take note that if u is set "too little" (θ ≈ 0), each film in the preparation test will be weighted generally the same, in this manner diminishing biased power. Interestingly, if u is too huge contrasted with the scope of d ($\overline{\overrightarrow{x_\iota}},\overrightarrow{x_l}$), the weighted normal in Eq. (5) will be commanded by the nearest tantamount, in this way expanding the inclination of our expectation. Given the applied contention above, we set the esteem of θ by speaking to studios' space information In this way, we select u with the end goal that any "comp" past the tenth will get insignificant weight; this is accomplished by setting u so that, by and large, in this way the eleventh or further comps have weights that are insignificant. In particular, we set u to such an extent that exp(-θ($d_{(10)}$ -$d_{(1)}$) ≈¢(2)/¢(0), which result in θ =17^5

Calibration of the Feature Weight ($\vec{v}$)

Next, we adjust the "element weights" $\vec{v}$. As a beginning point, an apparently sensible "default" decision is to put measure up to weights on each factor, i.e., set v_j =1 for all j . We refer to this as the Kernel-I approach; we will assess its prescient execution versus the Kernel-II approach that includes unequal component weights. We propose the accompanying methodology, in view of regularization what's more, cross-approval, to adjust the component weights $\vec{v}$ for the Kernel-II approach. We initially characterize the forget one mean squared-mistake, LOOMSE, a key part of our target work depicted later, as takes after. We let i=1,…….n(n= 265) file the contents in the preparation test, also, let $\acute{z}_\iota$ (θ, k, $\vec{v}$)be the anticipated estimation of the (log-) box office income of the ith content, when all aside from the ith content are utilized as the preparation information. z_i indicates the genuine log-box office income for the ith content. The LOOMSE is characterized as

$$\text{LOOMSE}\,(\theta, k, \vec{v}\,)=\frac{1}{n}\sum_{i=1}^{n}\left(z_i-(\theta,k,\vec{v}\,)\right)^2 \quad \text{……………(6)}$$

While it is enticing to straightforwardly limit LOOMSE as a capacity of $\vec{v}$, past research [10] has demonstrated that such approach tends to prompt over-fitting. In this way, limits the components of $\vec{v}$ to take just a limited arrangement of qualities, accordingly lessening the degrees of opportunity in$\vec{v}$. we propose utilizing a "regularization" approach to abstain from over-fitting. In particular, notwithstanding LOOMSE, we include a "sentence term" that punishes the separation between $\vec{v}$and the vector of 1 s (which compare to the from the earlier supposition that all highlights are weighted similarly)

$$\text{Objective Function} = \text{LOOMSE} + \lambda \sum_{j=1}^{J}(v_j{}^2 - 1)^2 \text{……..(7)}$$

We at that point adjust (the degree of the many-sided quality sentence) utilizing a cross-approval approach (see Appendix II, accessible on the web), which brings about a decision of λ= 0:05. At long last, given the decision of , we limit the target work in Eq. (7) as a component of utilizing the

Nelder-Mead technique , to touch base at the Kernel-II approach. As can be seen in Table3 the five most critical highlights are (I) early composition, (ii) generation spending plan, (iii) solid adversary, (iv) sort sentiment, and (v) classification back chiller. Further, master scriptwriters [6], likewise trust that a solid adversary is key for progressing the story line, as it helps set up a contention between the hero what's more, his/her adversary.

F) SVM: Support Vector Machine (SVM)

Support vector machines (SVMs) are one of the discriminative characterization techniques which are generally perceived to be more exact. The SVM characterization strategy is exceptional from the others with its exceptional characterization viability Besides, it can handle reports with high-dimensional info space, and separates out a large portion of the superfluous highlights. So SVM is regulated learning strategy for arrangement to discover the straight isolating hyper-plane which expand the edge, i.e., the ideal isolating hyper-plane (OSH) and augments the edge between the two information sets. To ascertain the edge, two parallel hyper planes are developed, one on each side of the isolating hyper-plane, which are "pushed up against" the two informational indexes.

Table 2 Calibrated Feature Weights in the Kernel-II Approach and S.V.M.

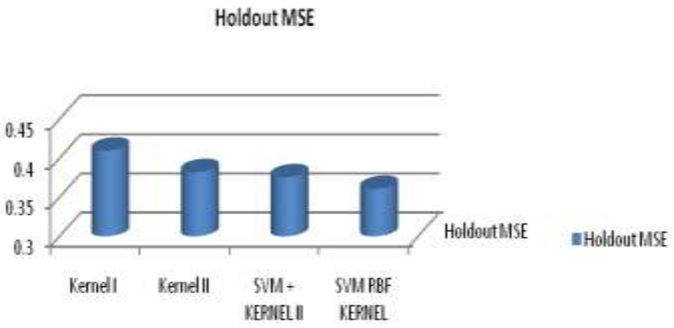
Variable	v_j^2 (Kernel II)	v_j^2 S.V.M.
EAREXP	**1932.9750**	**1816.9965**
BUDGET	**1902.0749**	**1825.9919**
STRNEM	**1699.50735**	**1631.5270**
GENRE_ROM	**1548.4400**	**1486.5024**
GENERE_THR	**1455.7396**	**1397.5100**
LS2	**981.93758**	**942.66008**
ANTICI	**923.57066**	**923.57066**
GENRE_SCI	**676.36959**	**649.31481**
FAMSET	**672.9362**	**646.01879**
RESOLUT	**593.9692**	**570.21046**
CHARGROW	**405.1350**	**388.92968**
LS1	**401.7017**	**385.63366**
FLHBACK	**398.2683**	**382.33765**
GENRE_ACT	**398.2683**	**382.33765**
MULCONF	**398.2683**	**382.33765**
DIAGCONC	**381.1016**	**365.85758**
GENRE_COM	**377.6683**	**362.56156**
COAVOID	**357.0682**	**342.78548**
INTCONN	**353.6348**	**339.48946**
LOGIC	**353.6348**	**339.48946**
LOCKIN	**346.7681**	**332.89744**
BUILD	**346.7681**	**332.89744**
HEROW	**343.3348**	**329.60142**
INTPERC	**339.9014**	**326.30541**
GENRE_FAM	**336.4681**	**323.00939**
INTENSITY	**336.4681**	**323.00939**
BELIEVE	**329.6014**	**316.41736**
IMP	**329.6014**	**316.41736**
SYMHERO	**326.1680**	**313.12135**
SURPEND	**326.1680**	**313.12135**
IMPPREM	**322.7347**	**309.82534**
MULDIM	**322.7347**	**309.82534**
SURP	**298.7012**	**286.75324**
CLRMOT	**288.4012**	**276.8651**
GENRE_DRA	**271.2345**	**260.3851**
CLRPREM	**264.3678**	**253.7930**
AVGDIAGLEN	**264.3678**	**253.7930**
GENRE_HOR	**264.3678**	**253.7930**

IV EXPERIMENTAL RESULTS

Keeping in mind the caveats of our data set the proposed Kernel-I/II approaches to predict the box office revenue of each movie in the holdout sample for example (movies released after January 2009). Predictive performance is measured by mean squared error on log- box office revenue. Remembering the provisos of our informational index, we apply the proposed Kernel-I/II , SVM with RBF and kernel II approaches to anticipate the movies income of every film in the holdout test (35 movies discharged after January 2009). The Kernel-I approach which uses all textual variables (with equal weights) improve holdout predictive performance even further with a holdout MSE 0.4096. Kernel-II approach which allows for unequal feature weights has the lowest MSE across all methods 0.3822. SVM with Kernel-II holdout MSE 0.3756 and SVM RBF Kernel holdout MSE 0.3607.

Table 3: Holdout Prediction Performance on 35 Movies Released After 2009

Method	Holdout MSE
Kernel I	**0.4096**
Kernel II	**0.3822**
SVM + KERNEL II	**0.3756**
SVM RBF KERNEL	**0.3607**

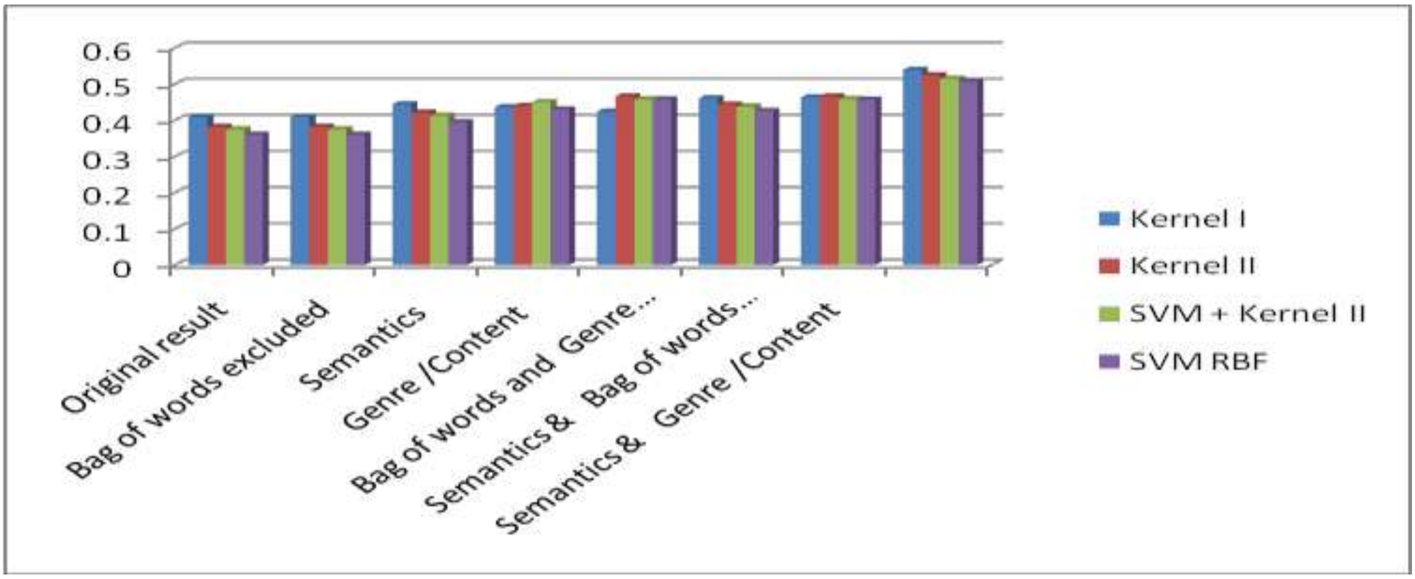


Graph 1: Holdout Prediction Performance on 35 Movies Released After 2009

The final result is shown in Table 4 Predictive Performance (in Terms of MSE) for Kernel-I and Kernel-II, SVM with Kernel-II and SVM with RBF kernel Methods, When Different Sets of Predictors Are Excluded.

Table 4 Predictive Performance (in Terms of MSE) for Kernel-I and Kernel-II Methods, When Different Sets of Predictors Are Excluded

	Kernel I	Kernel II	SVM + Kernel II	SVM RBF
Original result	0.4096	0.3822	0.3756	0.3607
Bag of words excluded	0.4096	0.3822	0.3756	0.3607
Semantics	0.4446	0.4219	0.4129	0.3950
Genre /Content	0.4362	0.4400	0.4507	0.4304
Bag of words and Genre /Content	0.4238	0.4665	0.4580	0.4586
Semantics & Bag of words and Genre	0.4613	0.4433	0.4378	0.4258
Semantics & Genre /Content	0.4639	0.4665	0.4589	0.4573
Genre /Content & Bag of words and Genre /Content	0.5396	0.5250	0.5150	0.5075



Graph 2: Predictive Performance (in Terms of MSE) for Kernel-I and Kernel-II Methods, When Different Sets of Predictors Are Excluded

V CONCLUSION

In this paper, we built up a system, in view of the piece based approach, to foresee the income capability of motion picture contents for the purpose of green-lighting. We gathered a database of 35 films contents and extricated three layers of printed data from each content: class/content, semantics, also, sack of-words factors, utilizing a blend of screenwriting space information, human info, and characteristic dialect handling strategies. Holdout forecast comes about recommended that our proposed Kernel-I, Kernel-II, SVM with RBF Kernel and SVM Kernel-II approaches beat relapse, tree-based strategy, and the comps-based approach our proposed approach talks straightforwardly to the instincts of studio supervisors, prompting key correspondence focal points which additionally improve the significance of our approach. Counting such highlights may help enhance prescient execution significantly further. A few difficulties remain, including the most effective method to scale up the database and refresh/keep up it over time, and how to convey our discoveries to studios most successfully.

REFERENCES

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

[2] X. Yu, Y. Liu, J.X. Huang, and A. An, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain," IEEE Trans. Knowledge and Data Eng., Vol. 24, No. 4,pp. 720-734, Apr. 2012.

[3] H. Chipman, E. Geroge, and R. McCulloch, "BART: Bayesian Additive Regresion Trees," The Annals of Applied Statistics, vol. 4, no. 1, pp. 266-298, 2010.

[4]http://esatjournals.net/ijret/2013v02/i12/IJRET20130212009 Online Review Mining For Forecasting Sales

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation,"J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[6] S. Field, Screenplay: The Foundations of Screenwriting. third ed., Dell Publishing, 1994.

[7] J. Monaco, How to Read a Film. Oxford University Press, 2000.

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

[9]M.F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14,no. 3, pp. 130-137, 1980

[10] G.J. Alexander and M.B. Alexandre, "Economic Implications of Using a Mean-VaR model for Portfolio Selection: A Comparison with Mean-Variance Analysis," J. Economic Dynamics and Control, vol. 26, no. 7/8, pp. 1159-1193, 2001.

[11] H. Mukerjee, "Nearest Neighbor Regression with Heavy-Tailed Errors," The Annals of Statistics, vol. 21, no. 2, pp. 681-693, 1993.

[12] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," J. Am. Soc. for Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[13] http://www.imdb.com

[14] http://www.gomolo.com