# Predicting Movie Success from Tweets

Rose Catherine
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
rosecatherinek@cs.cmu.edu

Sneha Chaudhari
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
sschaudh@andrew.cmu.edu

## 1 Introduction

Entertainment industry is one of the biggest industries in terms of expenditure, revenue and number of people involved, and has been growing steadily over the past few decades. Therefore, it is unsurprising that predicting whether a movie will be a great commercial success or not, generates considerable interest in both commercial and research communities.

With the advent of internet and social media, the behavior of average consumers is no longer an isolated action. The "wisdom of the crowd" is increasingly playing an important role in all matters ranging from the brand and model of the products that users buy, to the locales and resorts to holiday in; or in this case, deciding whether to watch a movie or not. Previous studies have shown that the hype created on social media prior to the release of a movie has a huge impact on its opening weekend success. Notwithstanding the research that shows a positive correlation between tweets and movie performance, tweet data is not always very reliable. Tweets are short text segments frequently with non-standard acronyms and abbreviations, leading to sparse and noisy representation of the data, which adversely affects the prediction accuracy.

In contrast, movie reviews present a less noisy and more descriptive version of the same content. Even though a more accurate prediction can be obtained from reviews, since their availability prior to the release of the movie is limited[1], their utility in predicting the performance of an unreleased movie is minimal. This opens up a middle ground which has not been explored before: using reviews to learn accurate predictors and transferring that knowledge onto the less predictable tweets which are available prior to the release, using the technique of Transfer Learning.

The aim of this paper is to investigate the possibility of accurately predicting the box office success of upcoming movies using tweets. The main contribution of this work is in demonstrating how this can be achieved using the technique of transfer learning. To the best of our knowledge, this is the first work that explores this golden mean for the task of box-office success predictions.

## 2 Related Work

Since the founding of Twitter in 2006, researchers have applied the collective wisdom of the twitterati to a wide array of research problems. In [2], the authors built a simple linear regression model that used the rate of tweets generated prior to the release of the movie to predict its gross revenue. [7] and [10] investigated how favorability and visibility of the movie on social media impacted movie sales and found that the popularity of the actors as reflected by their follower count in Twitter was one of the factors that correlated strongly with the success of the movie. This effect was also explored by [9] and [1]. Although it is obvious that having a large follower count on Twitter does not always guarantee that all movies of that actor or director will be a hit, it does shed some light onto the importance of Twitter as a medium which has the potential to reach masses as well as convert them into dollars.

In the case of movie reviews, [6] showed that the reviews correlate with the late and gross box office revenues, but do not exhibit significant correlation with early box office receipts. This is reasonable given that the movie reviews are usually available only after it has been released. In [8] and [3], authors used the review data to predict the revenues.

## 3 Proposed Approach

In many data mining and machine learning applications, it is often observed that there is a classification task in one domain of interest (target domain), but lacks sufficient data to train a decent model. However, there exists copious amounts of training data in another domain (source domain). In such applications, it is often feasible to use the knowledge from the source domain to improve the performance of learning on target domain by avoiding the much expensive data-labeling efforts.

In the setting of this paper, the target domain consists of data from Twitter, which includes tweets of upcoming / unreleased movies and the source domain comprises of the movie reviews of the released movies. Binary labels for released movies are available, which indicate whether the movie was a box office success or not. Hence, in this setting, we have unlabeled data in the target domain and sufficient labeled data in the source domain, with the task in both the domains being the same, which is a classification task. As a consequence, the underlying problem is a category of transfer learning called "Transductive Transfer Learning", in which the target domain has no labeled examples but the source domain has abundant labeled examples available. Further, it can be noticed that, the feature spaces in the source and target domain are different as the vocabularies and their sizes will be different for the reviews and the tweets. Consequently, the approach to transfer the knowl-

[1]Occasionally, reviews appear in Rotten Tomatoes prior to the release of the movie, usually because of pre-screening to a limited audience. However, such reviews are relatively few.

edge from source domain to target domain is called "feature-representation-transfer" approach, where the aim is to find a "good" feature representation which can minimize domain divergence and hence the classification error.

## 3.1 Technique

Given the data from the source and the target domains, the goal is to find a transformation mechanism that converts both datasets into a common reduced feature space, where the words of the source domain and their corresponding words in the target domain (which may have a different morphological form), both get mapped to the same feature in the common reduced feature space. The technique used in this paper is similar to the approach used by Blitzer et al. in [5], where they apply Transfer Learning to part of speech tagging, for the purpose of domain adaptation.

Let $X^R \in n^R \times t$ and $X^T \in n^T \times t$ be the doc-term matrices corresponding to training data from reviews and the Twitter domain, respectively. Here, $n^R$ is the number of review documents and $n^T$ is that of twitter documents. $t$ is $|M^R \cup M^T|$, where $M^R$ is the set of words in reviews, and $M^T$, that of twitter. i.e. the words are assigned ids from a joint space. The transformation mechanism uses a transformation matrix $\theta$, which is computed on the joint representation of the documents, $X = \begin{bmatrix} X^R \\ X^T \end{bmatrix}$. The algorithm for predicting the labels of the movies is given in Algorithm 1, where it first computes $\theta$ on the full set of training documents (Step 3). Then, it learns a classifier on the $\theta$-transformed version of the reviews (Step 4). In the testing step, it first transforms tweets using $\theta$ and then uses the above classifier to predict their success (Step 5).

The steps for computing $\theta$ is outlined in Algorithm 2. Note that the computation of $\theta$ does not require the labels of the documents, and thus, is completely unsupervised. Learning $\theta$ relies on what is called "Pivot" features, which are words that behave in the same way in both the domains, and hence, can be used to find the correlation between non-pivot features across both domains. Minimum requirements on pivot features are that they should occur frequently in the two domains and should behave similarly in both. There are different methods to determine pivots, like using Mutual Information [4]. But using the most frequent words have been shown to perform well enough [5]. In this paper, we use the top $m$ most frequent words of each of the domains as pivots.

In Algorithm 2, $Pivots$ are the indices of the pivot words, $P^R \cup P^T$ of the reviews and the twitter data; $|Pivots| \leq 2m$ i.e. pivots from the two domains may overlap. For each pivot $p$, we build a binary classification problem of predicting whether that pivot word is in the document or not. Thus, the ground truth for this classification is the $p^{th}$ column of $X$ (Step 3). The goal of this binary classification problem is to discover other words, obtained by removing the $p^{th}$ column of $X$ (Step 4), in the corpus that are predictive of $p$. For this reason, the loss function $L$ has to be a function of the inner product of feature vector and the weight vector. We use Logistic Regression which uses the Logistic loss function, to compute the weight vector. This weight vector (Step 5) essentially encodes the predictive power of other words with respect to $p$. Finally, all the $|Pivots|$ weight vectors $\hat{w}_p$ are appended together (Step 6) to construct a $W$ matrix of $t \times |Pivots|$. To compute a low-dimensional transformation matrix $\theta$, compute the Singular Value Decomposition of $W$ to get $[U\ \Sigma\ V^\top]$. The transformation matrix $\theta$ is the first $h$ columns of $U$ (Step 8), which are the left singular vectors corresponding to the $h$ largest singular values of $W$. And it transforms any document in the original $t$ feature space to a reduced common feature space of $h$ dimensions.

Given $\theta$, the new training data for the movie success prediction problem is the transformed doc-term matrix $X^R \cdot \theta$, which gives equivalent representation in the $h$D common feature space. A classifier is learned on this new training data, which is subsequently applied to the new test data given by the transformed doc-term matrix $X^{T'} \cdot \theta$, which is the equivalent representation of the twitter data in the $h$D common feature space.

**Algorithm 1** Predict Success

1: **procedure** MAIN($X^R$, $y^R$, $X^T$, $Pivots$, $X^{T'}$)
2: $X = \begin{bmatrix} X^R \\ X^T \end{bmatrix}$
3: $\theta =$ COMPUTETHETA($X$, $Pivots$)
4: $C = predictor(X^R \cdot \theta, y^R)$
5: **return** $C(X^{T'} \cdot \theta)$

**Algorithm 2** Compute Transformation Matrix $\theta$

1: **procedure** COMPUTETHETA($X$, $Pivots$)
2: **for all** $p$ **in** $Pivots$ **do**
3: $y_p = X_{col:p}$
4: $X_p = X \setminus X_{col:p}$
5: $\hat{w_p} = \arg\min_w L(X_p \cdot w, y_p)$
6: $W = [W|\hat{w_p}]$
7: $[U\ \Sigma\ V^\top] = SVD(W)$
8: $\theta = U_{col:1-h}$
9: **return** $\theta$

# 4 Experiments and Results

We shortlisted 101 movies from IMDb that have at least 500 reviews and for which we could obtain at least 100 tweets that appeared before the release date of that movie. Each reported result is an average of 10 random trials of the corresponding experiment, which in turn is a 3-fold cross validation on the corresponding data. To determine the ground truth labels, the budget and gross revenue of all the movies were collected from IMDb. Movies that made a profit of at least 30% of their budgets were labeled as Hits (positive) and the rest were labeled as Flops (negative). This gave us 42 positive and 59 negative examples.

## 4.1 Experimental Evaluation

The Transfer Learning framework has two parameters that need to be tuned — number of pivots, $m$ and the number of features after transformation, $h$. Due to space constraints, we provide the plot of only tuning $h$. $m$ was tuned to 20. Figure 1 gives the plot of F measure[2] with the classifier trained after knowledge transfer while varying $h$ from 100 to 1000, along with the classifier that was trained on just the Twitter data. Both correspond to using the most frequent 1000 terms from the raw doc-term matrix. Since the classifier on just the Twitter data does not have parameters $m$ and

[2]https://en.wikipedia.org/wiki/F1_score

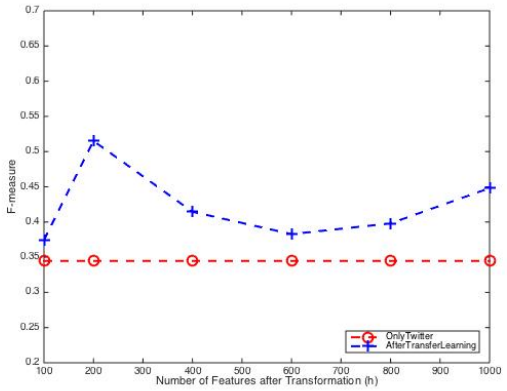


**Figure 1: Plot of F measure on `Dataset-Coarse` with and without Transfer Learning, and varying $h$ with $m = 20$**

$h$ it does not vary in the plot. From the figure, it can be observed that the classifier after transfer learning performs better than on the one with just the Twitter data. The best F measure obtained is 0.52 with $h = 200$, which is ∼53% higher (statistically significant) than that achieved by the baseline (0.34).

This demonstrates the efficacy of the proposed approach in successfully improving the classification performance on Twitter data for the task of predicting the box-office success of movies, using the technique of transfer learning.

## 5 Conclusions

Forecasting the box office success of upcoming movies is an important task for the entertainment industry, and is inherently complex due to its extremely unpredictable nature. Prior work has used Twitter data analysis to predict the success. However, the noisy nature of Twitter texts rendered them unreliable. In this work, we proposed a transfer learning approach to overcome the issues related to tweets. To accomplish this, we made use of the much cleaner and more descriptive online reviews of already released movies. In the proposed approach, we performed knowledge transfer by learning a common feature representation that helped to reduce the divergence between the reviews and the Twitter data. This enabled us to train a more accurate classifier on the transformed reviews data and employ it for predicting on the transformed tweets.

We presented results that compared the performance of classifiers trained on Twitter data alone, as well as the results obtained after performing transfer learning on Twitter data using the proposed approach, which improved upon the baseline by as much as 53%. The statistically significant results unequivocally point to the fact that it is possible to obtain highly accurate predictions from noisy tweets by leveraging information from the cleaner reviews data.

## 6 References

[1] K. Apala, M. Jose, S. Motnam, C.-C. Chan, K. Liszka, and F. de Gregorio. Prediction of movies box office performance using social media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 1209–1214, Aug 2013.

[2] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[3] S. Basuroy, S. Chatterjee, and S. A. Ravid. How critical are critical reviews? the box office effects of film critics, star power, and budgets. In *Journal of Marketing*, volume 67, pages 103–117, 2003.

[4] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447. Association for Computational Linguistics, 2007.

[5] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128. Association for Computational Linguistics, 2006.

[6] J. Eliashberg and S. M. Shugan. Film critics: Influencers or predictors?. *Journal of Marketing*, 61(2):68, 1997.

[7] J. Huang, W. F. Boh, and K. H. Goh. From a social influence perspective: The impact of social media on movie sales. 2011.

[8] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 293–296, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[9] J. J. Kaplan. Turning followers into dollars: The impact of social media on a movieâĂŹs financial performance. In *Undergraduate Economic Review*, volume 9, 2012.

[10] Shruti, S. Roy, and W. Zeng. Influence of social media on performance of movies. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2014.