The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Exploiting Textual, Visual, and Product Features for Predicting the Likeability of Movies

Mahsa Shafaei
University of Houston
Houston,TX
mshafaei@uh.edu

A. Pastor López-Monroy
Mathematics Research Center (CIMAT)
GTO, Mexico
pastor.lopez@cimat.mx

Thamar Solorio
University of Houston
Houston,TX
solorio@cs.uh.edu

Abstract

Watching movies is one of the most popular entertainments among people. Every year, a huge amount of money goes to the movie industry to release movies to the market. In this paper, we propose a multimodal model to predict the likability of movies using textual, visual and product features. With the help of these features, we capture different aspects of movies and feed them as inputs to binary and multi-class classification and regression models to predict IMDB rating of movies at early steps of production. We also propose our own dataset consisting of about 15000 movie subtitles along with their metadata and poster images. We achieve 76% and 63% weighted F1-score for binary and multiclass classification respectively, and 0.7 mean square error for the regression model. Using prediction methods and data analysis, this research helps the movie business to be more productive.

Introduction

Over the years, the number of released movies has increased massively (Dodd 2016), but according to Internet Movie database (IMDB)[1], only a few out of millions of movies get a high rating (higher than 8). As making movies is expensive, predicting likability of movies can significantly affect the movie industry. For example, movies like "Jupiter Ascending" and "The Lone Ranger" spent millions of dollars on production, but their IMDB rating is less than 6.5 (which shows these movies are not very popular), and also they could not make a profit in movie theaters. So, movie investors may lose a great amount of money by working on movies that are not liked by people. The cost of movie production comes from different sources such as production, marketing, screenings and financing costs. Our proposed method can be used as a tool by movie production companies (e.g., Pixar, Walt Disney, and Sony) to avoid most of these costs by early success prediction.

In this paper, our goal is to automatically predict the IMDB rating for the movies as a likability criterion. There are several works that introduce "Box Office Gross" as a success criterion, and they tried to predict this value for the movies (gross value shows the amount of money that movie earned from the box office). But, as we mentioned earlier, our criterion is the IMDB rating not gross revenue because of four main reasons. First, unlike the IMDB rating, the gross revenue is not available for a large number of movies. Second, the price of box-office ticket changes during the years, so we cannot compare old movies with newer ones. Third, the gross value depends on many other variables such as advertisements and competitor movies. Finally, movie theaters are not the only source for movies' revenue; there are other sources like home entertainment, television deals, and video on demand (i.e. Netflix and Amazon). Therefore, IMDB rating is a more reliable likability criterion.

Although intrinsic factors, such as quality of the screenplay and story of the movie, play an important role in the likability of movies, extrinsic factors including the popularity of directors and advertisement (e.g. movie posters) are equally important. In this research, we aim to exploit multimodal information by modelling textual, visual and production information. For this purpose, we extract textual features from movie subtitles, visual features from movie posters, and production features from movies' metadata to capture different aspects of movies. Using these features, we propose regression and classification models to automatically predict the IMDB rating for movies. Figure 1 shows the diagram of the whole system.

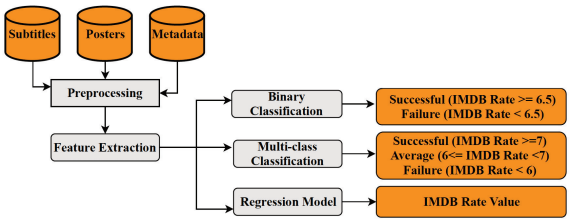

Figure 1: System diagram

It should be noted that we only use items that are available before movie screening. Although features released in later steps of the movie production (like movie awards) can

[1]IMDB is a well-known website for rating movies with 83 million registered users

help improve the prediction result, they cannot be helpful in a real scenario as companies or producers need to decide to start filming a movie or not at a very early stage. Our contributions in this work are as follows:

- Introducing a new dataset for movie subtitles along with their metadata. Our dataset is the largest one in movie success prediction field with about 15,000 movies.
- Defining a new set of semantic, syntactic and visual features that help us to achieve better performance in predicting IMDB rating (likability of movies) for both regression and classification models.
- Analyzing the correlation between the gross revenue and other potential factors like rating and genre.

Previous work

Predicting movies' likability can be solved by different approaches. These approaches are different in terms of likability definition and also timing of prediction. By timing, we mean that some of works use data that are available before the production, some of them use data that are available after production but before releasing, and some works propose methods that employ data even after the movie screening.

The first group are papers that predict gross revenue as a success criterion. Researchers in (Wagholi 2016), (Apala et al. 2013) and (Asur and Huberman 2010) used social media to predict movies' box office grosses. These works gathered data, like users' comments, from social media including Twitter and YouTube. They gathered comments that are written when the trailer of movies are released and the movie itself is not shown at the movie theater. So, they used data after production but before screening. Authors of (Lash and Zhao 2016) also tried to predict the revenue of movies, they considered the return on investment (Profit/Budget) as a success criterion. They defined both binary classification and multi-class classification and extracted three types of features: audience-based, release-based, and movie- based features. To extract features related to movies' concept, they used movies' plot synopses that are written by users and there is no standard structure for that.

The second group are researches that predict the likability of the movies based on IMDB rating. Authors of (Ericson and Grodman 2013) gathered around 4k movies and predicted binary success prediction based on IMDB rating (they considered 6.5 as a threshold), and they achieved 71% accuracy for the SVM classifier. Papers (Latif and Afzal 2016), (Asad, Ahmed, and Rahman 2012) and (Saraee et al. 2004) split movies into 4 classes according to IMDB rating (Terrible, Poor, Average, Excellent). Although (Latif and Afzal 2016) achieved a good accuracy, they used parameters that are available after movie screening like awards, number of screens, etc.

In this work, we predict IMDB rating for movies using our own dataset. In spite of previous works, we gather a dataset that consists of movie subtitles rather than movie scripts. Considering the fact that our dataset (with about 15k movies) is quite larger than similar works, we can claim our results are more reliable. To extract content-based features from subtitle, we borrow some features from (Ashok, Feng, and Choi 2013) and (Maharjan et al. 2017). They extracted lexical features, production rules, constituents, and sentiment features from book content to find the relation between books' writing style and their success. We combine these features with other features like visual and production features and improve the result compare to previous works that did not use data after screening.

Dataset

Despite movie transcripts, subtitles are available for a large number of movies and they have standard format. We collect subtitles from a freely available source (https:\\ www.springfieldspringfield.co.uk) and extract the text from HTML web-page. The other resource we use is "SubLight", which is an application for downloading the movie subtitles for free. The output of this application is in subtitle format and contains dialogue timing. So we delete extra data, and we keep conversations between characters.

To extract the metadata of movies, we employ IMDB API. With the help of this API, we download name of directors and actors, movies' genre, downloadable poster links, movies' run-time, etc. Using poster links, we also download poster images for all movies in the database.

Using all resources that we mentioned, we build a dataset that contains about 15,000 movie subtitles along with the metadata. [2] Despite the fact that IMDB provides valuable information about the movies, it is not the best resource for movies' box office grosses. Thus, we crawled websites "Box Office Mojo" and "The Numbers" to gather the box-office grosses. About 5,000 movies in our datast have this value.

Success Definition: As we mentioned before, we propose binary and multi-class classification and regression models. For classification models, we need to precisely define our definition for classes. For binary classification, same as (Ericson and Grodman 2013), we define the threshold 6.5 which means movies with a rating higher than 6.5 are successful while movies with a rating less than this value are not. In multi-class classification, we categorize movies into three groups. Movies with a rating higher than 7 are successful, movies with a rating between 6 and 7 are average and movies less than 6 are Unsuccessful. The reason behind this definition is that in this way, we have an approximately same distribution of data in each group. Table 1 reports statistics about dataset and Figure 2 shows the number of movies with a specific rating range. According to this plot, most of the movies have a rating between 5 and 8.

Multi-Class	#	Binary	#
Successful	5726	Successful	7551
Average	5486	Unsuccessful	8394
Unsuccessful	4733		
Total	15945	Total	15945

Table 1: Data statistics

Table 2 shows the distribution of data in each genre.

[2] http://ritual.uh.edu/1493-2/

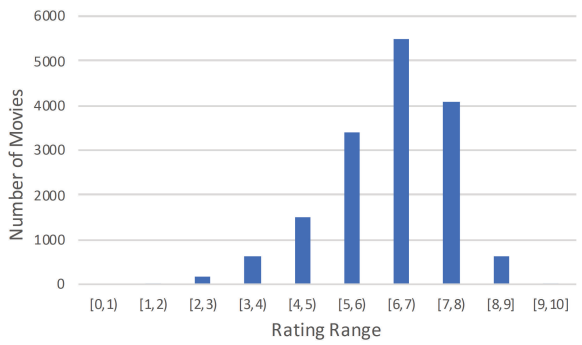

Figure 2: Movies distribution in each rating range

Genre	#	Genre	#
Science-Fiction	986	Action	2870
Horror	2024	Animation	663
Crime	2391	Adventure	1787
Romance	2607	History	540
News	22	Western	323
Comedy	4957	War	422
Thriller	2514	Short	324
Mystery	1096	Film-Noir	214
Musical	787	Drama	8043
Documentary	879	Family	934
Sport	304	Biography	833
Fantasy	862		

Table 2: Data distribution in each genre. Some movies are assigned to more than one genre, so sum of movies in all genres is higher than the total number of movies

Methodology

Our final goal in this work is to predict the likability of movies. We approach this problem by introducing two types of models; classification and regression. We use three sources as inputs in these models. We extract textual features related to lexical, semantic and syntactic aspects of subtitles. We also extract visual features from movie posters to capture important objects representing the movie. Moreover, there are some features related to movie production that are available at the early steps of movie production. You can see the feature diagram in Figure 3.

We group our features into two sections. Traditional features that have been used before for movie success prediction, and new features that we use for the first time in this field.

Traditional Features

Lexical We extract unigram and bigram features from subtitles and apply term frequency-inverse document frequency (TF-IDF) as the weighting scheme. In addition to basic ngram features, we extract two skip-ngram (n=2,3) from subtitles. K-skip-n-grams allows k or fewer skips to construct the n-gram.

Genre, Actors and Directors Previous works usually used these three elements of metadata to predict likability of

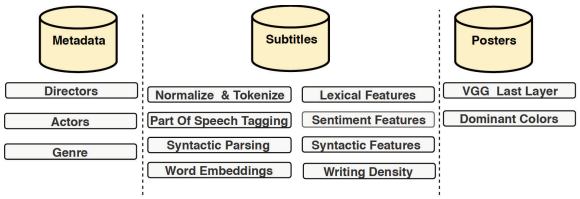

Figure 3: Feature Diagram

movies. In this work, we model them as binary vectors (binary bag-of-words). Moreover, website (https://www.ranker.com/) publishes a list of best actors according to the actors' popularity and achievements. We check if the first actor of a movie is in this list, we assign the corresponding score of the actor in the list as a feature value of the actor. And, if the movie's first actor is not on the list, we assign 0 to the feature.

New Features

Based on our knowledge, features in this section have not been used in movies' likability prediction before.

Sentiments We use three methods to extract sentiment in a text: NRC emotion lexicon , SenticNet, and SentiWordNet.

NRC emotion lexicon: Author of (Mohammad 2011) provides us with a dictionary of words mapped to eight different emotions (anger, anticipation, joy, trust, disgust, sadness, surprise, and fear) with binary values. To extract emotion flow, we divide the whole movie subtitle into n equal sections (n=5,10,15), count the words of each emotion, and return the values of all emotions for each section. As a result, we can have ups and downs of each emotion during subtitles of the movies (Kar et al. 2018). We also calculate the average of all emotions for the whole movie.

SenticNet: SenticNet provides a set of semantics, sentics, and polarity associated with 100,000 natural language concepts (Cambria 2013). Using the SenticNet parser, we extract sentiment concepts from the movie contents (Rajagopal et al.). To use these concepts in our model, we create binary bag-of-concepts features. To have a trend of emotions through the scripts, we divide the subtitle to n equal parts (same as what we did for NRC emotion lexicon), and we calculate the average of each emotion for each section (polarity, sensitivity, attention, pleasantness, and aptitude). We also calculate the average of all emotions for the whole movie as well.

SentiWordNet: SentiWordNet (Baccianella, Esuli, and Sebastiani 2010) provides positive and negative sentiment values for every synonym set. We use SentiWordNet same as (Maharjan et al. 2017).

Writing Density This feature is used by (Maharjan et al. 2017) for book success prediction. We also employ this feature to find out if a different density of elements like exclamations and question marks affect the quality of a movie script or not.

Word Embedding Word embedding is an effective technique for text classification because it is capable to capture

semantic information of the text. We calculate the average of all word vectors of every word in the subtitle and consider the value as a feature for our model. To do this, we use FastText [3] pre-trained word embedding.

Syntactic We use Stanford parser to extract parse trees for all sentences in the script, and we extract different production rules from parse trees. Lexicalized production, unlexicalized production, grandparent lexicalized production and grandparent unlexicalized production are production rules that we extract from parse trees. The idea is to capture the grammatical style of the subtitles.

Visual Features Typically, movie posters are pictures that demonstrate some important elements of a movie. So, we add some features related to posters into our model.

Last-layer output of VGG model: VGG model is one of the popular deep learning models for image classification. We apply the transfer learning method by initializing weights from the pre-trained model on ImagNet data. Then, we train our model with our dataset and use the output of the last layer before the fully connected layer of the model as a new feature for our classification/regression model.

Dominant color of posters Using an existing tool,[4] we extract three dominant colors of the posters. Then, we transform these colors to RGB codes and assign a number (0,1,2) to that color according to red, green and blue value dominance (the temperature of the color).

Experiments and Results

Our experiments are divided into three sections. First, we show the result of experiments for binary classification with threshold 6.5. Second, we report Multi-class classification results with three classes (with thresholds 6 and 7). Finally, as these threshold are subjective and there may be no consensus on them, we also build a regression model to predict the rating value regardless of any categorization and thresholds. The evaluation metric we use is weighted F-score for classification methods and mean squared error (MSE) for the regression model. To discover the effect of each feature on likability prediction, we run the experiments using each feature separately. We also do the experiments with different combinations of features to find the best combination of features for this task.

We start our task by pre-processing steps (e.g. converting all words to lowercase and removing infrequent tokens). Then, we extract features from data and randomly split data to 80:20 train and test sections. Finally, we train a Linear Support Vector Machine (Linear-SVM) classifier and linear regression model with extracted features.

According to the Table 3, the best mean squared error in regression model is **0.7**. To have an intuition about how good this result is, we compare it with a baseline method. To build the baseline model, we replace the value of predicted rating with "average rating" for all movies, and then calculate the MSE for the system. We consider this value as a baseline result. The MSE for the baseline is **1.44**, and it is twice the MSE of our model.

Features	BC(F1)	MC(F1)	MSE
Unigram(1)	0.72	0.50	10.11
Bigram(2)	0.73	0.56	1.36
2 skip 3 grams(3)	0.72	0.54	0.98
2 skip 2 grams(4)	0.73	0.56	0.94
NRC-emotion(5)	0.38	0.31	1.44
SentiWordNet(6)	0.55	0.31	1.79
SenticNet(7)	0.40	0.31	1.43
SenticNet-Concept(8)	0.61	0.31	1.32
Writing Density(9)	0.41	0.32	1.44
Syntactic(10)	0.68	0.47	1.95
FastText(11)	0.59	0.38	11.69
Genre(12)	0.67	0.54	1.07
Directors(13)	0.67	0.50	1.22
Actors(14)	0.65	0.48	11.44
VGG + DC(15)	0.53	0.43	3.19
1,2,3,4,12,13	**0.76**	**0.63**	**0.72**

Table 3: Results for classification (BC = Binary Classification, MC = Multi-class Classification) and regression tasks. MSE = Mean Squared Error, F1 score is weighted F1 scores, and DC stands for Dominant Color

Among all feature combinations, the best result in classification and regression methods is achieved by combining unigram, bigram, 2 skip 3 grams, 2 skip 2 grams, Genre and Directors. According to the results, lexical features (n-grams and skip n-grams) are one of the most important features in success prediction. These features are able to extract the pattern of word usage in the subtitles. Another important aspect of movies to be liked or not is the movie genre. So, based on the result some genres are more popular among people compared to others. Also, the combination of the genre with other features helps other features to be more effective. For example, some directors are more successful in a specific genre, so the combination of these two features makes an effective feature. Some features like actors and syntactic features produce a good result by themselves, but when we combine them with other features, they make no improvement on the overall results. So, we do not use them in the last version of our model to reduce the complexity. As you can see in the Table 3, emotion features are not strong features in movie success prediction task. It shows that there is no pattern for emotional ups and down in successful movies. In the next section, we do some in-depth analysis to have a better understanding of our data.

Data Analysis

In the experiment section, we presented the effect of different features on classification and regression models to predict movies' rating. Another important aspect of movies is how much a movie can earn in the movie theater. So, we analyze our data to find out if there is any correlation between movie revenue (or box-office gross) and IMDB rating. As we mentioned in the dataset section, we have box-office grosses for about 5,000 movies, so our analysis is only on this subset

[3] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

[4] https://github.com/algolia/color-extractor

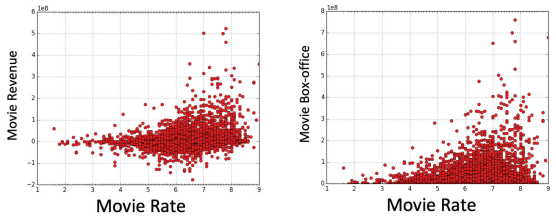

Figure 4: Right sub-figure shows relation between movie ratings and movie box-office values. Left sub-figure shows relation between movie ratings and movie revenue values (box office gross - budget)

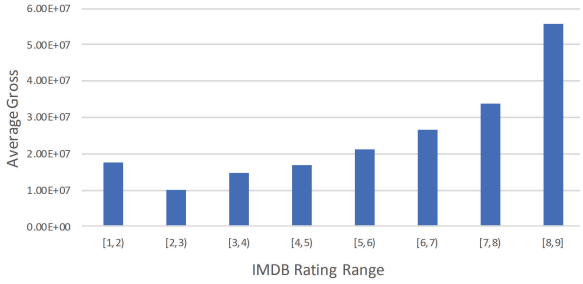

Figure 5: Average gross revenue for movies with a specific rating range

of the dataset.

In the right sub-figure of figure 4, we compare the trend of movie ranking and movie box-office. According to the results, although there is an increasing curve in the plot, there is no solid relation between these two parameters.

A high box-office gross cannot show if the movie has a high revenue or not because it is possible that a movie with a high box-office value also spent a high budget which means the final revenue of the movie is low. For instance "Pirates of the Caribbean: On Stranger Tides" earned about 240 million dollars from domestic movie theaters, but it also spent about 410 million dollars as budget. So, this movie lost money despite its high gross value. As a result, we also show the relation between revenue (Box office gross - Budget) and rating in the left sub-figure of Figure 4.

According to both sub-figure in 4, there are some movies that have a high or average rating but low revenue or box-office gross . On the other hand, all the movies with very low rating also earned very low money in the movie theaters. This is a reasonable outcome to have movies with high rating but low box-office gross because movies are shown on screens for a limited time, but they are available for rating during the years. Moreover, people all over the world can see and rate movies, but the gross revenue is limited to countries that show the movie in movie theaters (here we only used gross revenue at USA). Although we do not have a very high correlation between rating and box-office gross, Figure 5 shows that the average value of movies' gross increases by increasing the rate. In other words, the higher the rating, the higher the average gross revenue is.

We also separate movies in each genre and calculate the

Genre	Crr	Genre	Crr
Science-Fiction	**0.32**	Action	**0.32**
Horror	0.23	Animation	0.27
Crime	0.14	History	0.22
Romance	0.042	Adventure	**0.32**
News	-0.34	Western	**0.34**
Comedy	0.14	War	0.17
Thriller	0.25	Short	0.11
Mystery	0.16	Film-Noir	0.29
Musical	0.04	Drama	0.15
Documentary	-0.15	Family	0.2
Sport	0.13	Biography	0.23
Fantasy	**0.32**		

Table 4: Correlation between movie ranking and movie box-office gross in each genre

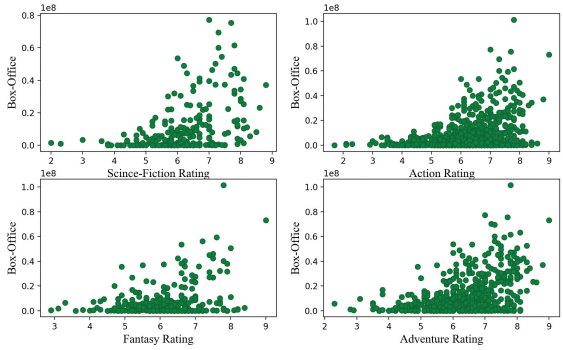

Figure 6: Each sub-plot belongs to a specific genre and shows the relation between rate and box-office gross of movies in that genre. X-Axis is rating, y-Axis is gross.

correlation between movies' rating and movies' box-office grosses for that specific genre. The goal of this experiment is to find out if there are some genres with high correlation between rating and box-office gross or not. According to Table 4, the five top correlations belong to Western, Science-fiction, Action, Adventure and Fantasy. It means that usually, people prefer to watch the high quality of movies in these genres at movie theaters. Figure 6 shows the correlation diagram for these five genres. On the other hand, there are some genres that cannot sell at movie theaters even though they are liked by people and have a high rating like Drama or Documentary. So, production companies can use this information to decide about their contract according to the media they want to show the movie on.

Another interesting aspect is the relation between movie genre and revenue or between movie genre and rating. We do these analyses to discover likability of movies in each genre in terms of ranking and selling at the box-office. According to top sub-figure in Figure 7, popular genres in movie theaters are Animation, Adventure and Science-fiction. Even though people enjoy to go to the cinema to watch these type of movies, getting high ranking is not very easy for movies in these genres. On the other hand, bottom sub-figure in figure 7 shows that the three top best ranking genres are Doc-

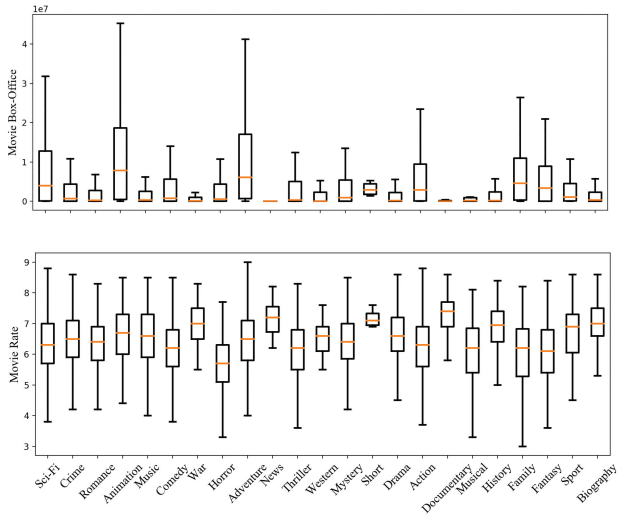

Figure 7: The x-Axis is name of different genres of the movies. Top sub-figure shows rating and bottom sub-figure shows box-office gross of movies in each genre.

umentary, Short-Films and News. This result can be interesting for companies like Netflix that do not care about box-office grosses and prefers to have shows with high rating and popularity.

Conclusions

In this paper, we proposed a method to predict likability (IMDB rating) of movies based on text-related, image-related and product-related features from movies. We presented a new dataset of 15k movie subtitles along with image posters and meta-data related to these movies. We achieved 0.7 mean squared error in the regression model, 0.76 and 0.63 weighted F1-score for binary and multi-class classification respectively. We also investigated the correlation between box-office gross and ranking, and we discovered that movie with high gross also have a high rating, but it is not true the other way around as there are movies with a high rating that did not earn high gross revenue. Finally, we determined genres with highest ranking and highest box-office gross.

Acknowledgments

We would like to thank the National Science Foundation for partially funding this work under award 1462141.

References

Apala, K. R.; Jose, M.; Motnam, S.; Chan, C.-C.; Liszka, K. J.; and de Gregorio, F. 2013. Prediction of movies box office performance using social media. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 1209–1214. ACM.

Asad, K. I.; Ahmed, T.; and Rahman, M. S. 2012. Movie popularity classification based on inherent movie attributes using c4. 5, part and correlation coefficient. In *ICIEV 2012*, 747–752. IEEE.

Ashok, V. G.; Feng, S.; and Choi, Y. 2013. Success with style: Using writing style to predict the success of novels. In *EMNLP 2013*, 1753–1764.

Asur, S., and Huberman, B. A. 2010. 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*, 492–499. IEEE Computer Society.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, 2200–2204.

Cambria, E. 2013. An introduction to concept-level sentiment analysis. In *Mexican International Conference on Artificial Intelligence*, 478–483. Springer.

Dodd, C. J. 2016. Theatrical market statistics. MPAA Washington, DC.

Ericson, J., and Grodman, J. 2013. A predictor for movie success. *CS229, Stanford University*.

Kar, S.; Maharjan, S.; López-Monroy, A. P.; and Solorio, T. 2018. MPST: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lash, M. T., and Zhao, K. 2016. Early predictions of movie success: the who, what, and when of profitability. *Journal of Management Information Systems* 33(3):874–903.

Latif, M. H., and Afzal, H. 2016. Prediction of movies popularity using machine learning techniques. *IJCSNS* 16(8):127.

Maharjan, S.; Arevalo, J.; Montes, M.; González, F. A.; and Solorio, T. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, 1217–1227.

Mohammad, S. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 105–114. Association for Computational Linguistics.

Rajagopal, D.; Cambria, E.; Olsher, D.; and Kwok, K. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *WWW 2013*, 565–570. ACM.

Saraee, M.; White, S.; Eccleston, J.; et al. 2004. A data mining approach to analysis and prediction of movie ratings. *Transactions of the Wessex Institute* 343–352.

Wagholi, P. 2016. Prediction of movies box office performance using social media. *International Engineering Research Journal*.