Screenplay Quality Assessment: Can We Predict Who Gets Nominated?

Ming-Chang Chiu Tiantian Feng Xiang Ren Shrikanth Narayanan
Department of Computer Science
University of Southern California
Los Angeles, CA 90089, USA
{mingchac,tiantiaf,xiangren}@usc.edu
shri@sipi.usc.edu

Abstract

Deciding which scripts to turn into movies is a costly and time-consuming process for filmmakers. Thus, building a tool to aid script selection, an initial phase in movie production, can be very beneficial. Toward that goal, in this work, we present a method to evaluate the quality of a screenplay based on linguistic cues. We address this in a two-fold approach: (1) we define the task as predicting nominations of scripts at major film awards with the hypothesis that the peer-recognized scripts should have a greater chance to succeed. (2) based on industry opinions and narratology, we extract and integrate domain-specific features into common classification techniques. We face two challenges (1) scripts are much longer than other document datasets (2) nominated scripts are limited and thus difficult to collect. However, with narratology-inspired modeling and domain features, our approach offers clear improvements over strong baselines. Our work provides a new approach for future work in screenplay analysis.

1 Introduction

The motion picture industry is a multi-billion dollar business worldwide (Lash and Zhao, 2016). Decisions in selecting movies to be produced are critical to the profitability of a movie studio. However, the selection of the screenplay that happens at the initial phase of a movie production pipeline and has a large influence on the financial budget and quality of the final movie production, has a large subjective element. For example, a typical script review service costs a studio $80 to $150 to receive a report containing a short summary of the script and opinion as to its quality (Follows et al., 2019). Considering the amount of scripts a studio needs to filter through, it can be overwhelming. Thus, an objective and reliable tool to help evaluate and narrow down the candidate scripts is of vital importance to aid the "green-lighting" (deciding which scripts to turn into movies) process.

Consider this scenario, if a tool can facilate the script review process and provide the chance of success, wouldn't this make an impact and cut down lots of budgeting decisions in the production process? The main idea of this work is to develop such a tool which gather custom analyses from various aspects, e.g., screenplay writing theory, character-focused linguistic behavior, to help assess the quality of the script.

In general, movie script writing can follow a well-defined *Three Act* structure (Field, 2007; McKee, 1997). Also, Weiland (Weiland, 2013, 2018) specifies a more fine-grained storytelling plan, starting from *hook, inciting event, 1st plot point, 1st pinch point, midpoint, 2nd pinch point, 3rd plot point, climax* to *resolution*, what are called Structural Points (SP). We believe knowledge like the above in strucuring a screenplay can bring benefits in selecting the most relevant textual properties for the prediction of script quality.

Aside from the event positioning, Follows et al. (2019) reported that how writers develop characters and events, i.e., *Characterization* and *Plot*, are two main foci of industry reviewers. We thus devise our domain specific features in these two aspects. We hope to offer an enhanced understanding of the essential elements in high-quality movie scripts.

To perform quality assessment, based on an assumption that the nominated scripts are recognized writings and thus should have had higher chance of passing green-lighting, we propose to perform an evaluation in a two-fold approach. First, we use award-nomination prediction as a proxy to the green-lighting process. Second, we examine our domain features and models by integrating them into existing document classification methods.

We acknowledge the constraints of our metric in that the number of award venues has its limits, and

Proceedings of the 1st Joint Workshop on Narrative Understanding, Storylines, and Events, pages 11–16

not necessarily those without nomination would be any worse than the nominated. But due to the difficulty in collecting unproduced scripts with peer reviews, we adopt our current approach.

Our main contributions are as follows: (1) We defined a quality metric for screenplays and collected ground truths from peer-reviewed venues. (2) Based on structural knowledge of screenplay narratology, we developed a simple narratology-inspired model for our task. (3) Motivated by industry opinions and narratology, we devised domain-specific features to achieve our objective. (4) We tested that for long document classification, a simple feature-based approach can work better than state-of-the-art models.

2 Related Works

Literary works-related research has gained interest in recent years. Bamman et al. (2013, 2014) have succeeded to learn latent character types in film and novels; Iyyer et al. (2016); Chaturvedi et al. (2016); Elson et al. (2010) try to model character relations in novels. Papalampidi et al. (2019) analyze narrative structure of movies by using turning points, and Chambers and Jurafsky (2008); Sims et al. (2019) seek to detect events in narratives. On text quality assessment, Mesgar and Strube (2018) encode local change patterns to assess readability and score essays; Toledo et al. (2019) collect argument pairs that was originally built for an automatic quality assessor for debate.

A noteworthy attempt in measuring quality of literary works we know of is made by Kao and Jurafsky (2012), who quantitatively analyze various indicators for discerning professional poems from amateurs'. However, in script writing, the cinematic success criteria lack evaluative consensus (Simonton, 2009) — previous works on evaluation of movies have largely focused on forecasting revenue or profit of movies using production, distribution, and advertising data (Ghiassi et al., 2015; Lash et al., 2015) or basic textual and human annotated features (Eliashberg et al., 2014).

The main differences between our work and previous works are: (1) our approach aims to process automatically without human annotated features. (2) our metrics and methods are geared towards evaluation that based solely on textual properties.

3 Data and Problem Setting

Data collection. We evaluated our method using ScriptBase (Gorinski and Lapata, 2018) and Movie Screenplay Corpus (MSC) Ramakrishna et al. (2017) datasets. ScriptBase provides 917 scripts and MSC contains 945 Hollywood movies. We kept 897 and 868 suitable ones which have enough character utterances for our approach from each dataset respectively. Similar to Underwood (2019), which analyzes high-prestige novels as works that have been reviewed by top journals, we collected the screenplays that have histories of nominations as quality "ground truth". The venues we collect from are well-known professional prizes, which include "Writers Guild of America Award", "Academy Awards", "Golden Globe Awards", and "British Academy of Film and Television Arts Awards". We assume the nominated scripts are of higher quality by professional standards. Since we focus on textual properties for success, we only gleaned nominations in the "original screenplay" and "adapted screenplay" categories. In the end, we obtained 212 (23.6%) movies out of ScriptBase and 113 (13.0%) from MSC as quality "ground truth" labels.

Problem Setup. Our work focuses on measuring quality as whether or not a movie would be nominated at a peer-reviewed venue. The basic assumption for using this approach as success metrics is simple — a screenplay that receives nominations by critical reviewers should have had higher chance of getting through green-lighting.

Challenges. By nature, a movie should be tough to be cleanly categorized, due to its length, complex storyline and turns, and the lack of evaluative criteria. Prior works in document classification (Yang et al., 2016; Liu et al., 2017; Adhikari et al., 2019; Johnson and Zhang, 2015) evaluated on datasets with small document size (Reuters, IMDB, Yelp, etc.). However, our document size on average is at least 65 times longer, which may be challenging for NN-based models to train due to long sequences and the associated computational burden. Besides, the number of training data we have is at most 1000 times smaller than other datasets. With our datasets being **long**, **fewer** and **skewed**, state-of-the-art deep learning techniques may not work well. Summary of the comparisons is shown in Table 1.

Dataset	documents	average #w	%pos
Reuters	10,789	144.3	-
IMDB	135,669	393.8	-
Yelp 2014	1,125,457	148.8	-
ScriptBase	897	27,539.7	23.6
MSC	868	27,067.4	13.0

Table 1: **Dataset statistics and comparisons of datasets.** #w denotes the number of words and %pos denotes the percentage of positive class.

4 Analysis of Domain Features

In this section, we introduce our domain features that are divised to achieve our goal and provide analysis based on our problem setup.

Characterization and *Plot* are major aspects of focus in the industry; inspired by which, we devised 6 novel features. For each, we provide intuitive motivations, and then detail how we converted them computationally. We chose the top two most speaking characters of each movie to analyze for *characterization*.

According to Weiland (2018), a script can place 9 SPs roughly equally distributed, creating eight equal-lengthed development segments (DS) in between. We hypothesize that such structural hints should help to achieve our objective. Based on the statistics of both datasets, to leverage the SPs, we collected a context window of 1% (∼270 words) centered at SPs for all scripts. Larger windows may contain more information and should improve the results, and we leave that for future experiments.

By the definition of characterization, we hypothesized that by measuring pattern change of characters, we may see how writers develop the characters' personality. We sought pattern change via two kinds of changes writers would make between SPs - linguistic (speaking pattern) change and emotional change. To do this, we proposed *Linguistic & Emotional Activity Curve*.

Linguistic & Emotional Activity Curve (*ling, emo*). For linguistic change, we extracted the dependency trees of characters; for emotional change we used normalized Empath (Fast et al., 2016) to get characters' emotion status. We combined the linguistic distribution, Empath distribution of sentences in each DS with *activity curve* (Dawadi et al., 2016), which uses a Permutation-based Change Detection in Activity Routine (PCAR) algorithm, to measure the change between two DSs of distributions.

Type-token ratio (*tt*). As Kao and Jurafsky (2012) show, in poetry, the *type-token ratio* related most positively to the quality of a poem. We believed this concept should work similarly on character analysis, and can show how much effort writers devoted in characterization. We defined this feature as the number of unique words used by a character divided by the total number of words.

Valence-Arousal-Dominance (*VAD*). Mohammad (2018a) performed extensive study in getting an objective score for words in VAD dimensional space (Russell, 1980, 2003). We used average scores over the context window of each SP to represent level of emotion.

Emotion Intensity (*int*). Similar to *VAD*, we used the NRC Affect Intensity Lexicon (Mohammad, 2018b) over the SPs to score emotion intensity along four basic emotion classes (Plutchik, 1980).

Also, since events are usually addressed in units of scenes, we wanted to get a picture of how many different emotionally similar scenes across the dataset appear in a movie.

Empath Clustering (*clus*). We retrieved lexical categories for each uttrance from Empath and then clustered the lexical category distributions of all utterances with deep embedded clustering (Xie et al., 2016). We obtained the cluster distribution based on the lexical categories within a movie as a feature representation.

We visualized partial features in a "nomination vs non-nomination" fashion, as in Fig. 1, to show the potential of our features. For some we can easily observe clear differences from one to the other, while some are more subtle. For instance, in *VAD*, the *arousal* of MICA is ambiguous between the two, and yet we can easily discern nominated scripts along the same axis for ScirptBase.

5 Predictive Modeling

In this section, we define our prediction task, and then propose our base model and then move on to a paradigm which integrates domain features proposed in previous section.

Task Formulation. As a proxy to the original quality assessment task, we define a binary classification task as to predicting the nomination of a script.

Narratology-inspired Model. Inspired by narratology, we propose *Tfidf-SVM*$_{narr}$ — instead of using all texts in an entire document, we extract words in context window of SPs for each docu-

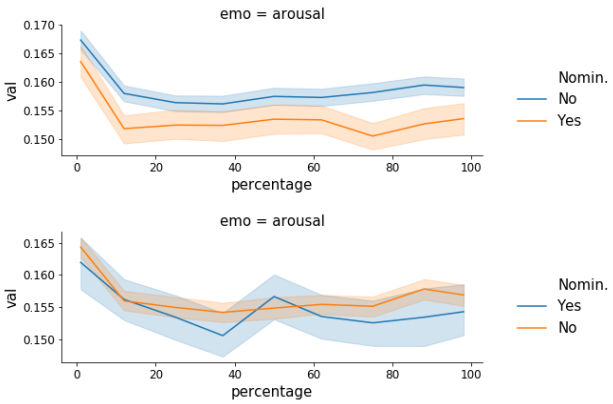

Figure 1: **Nomination vs Non-nomination of arousal level along percentage of scripts.** *Upper:* ScriptBase. *Lower:* MSC.

ment, compute the tf-idf representations, and feed them into a SVM classifier. The main components of *Tfidf-SVM*$_{narr}$ are shown in Fig. 2. Due to the large amount of unique tokens, we chose only the top 500 important features ranked by Tf-idf to represent a document. We test the results without choosing 500 features and our setting is better.

Feature-based Prediction. To examine the predictive power of proposed features, on top of *Tfidf-SVM*$_{narr}$, we add domain features along with tf-idf to SVM to see the efficacy of domain features.

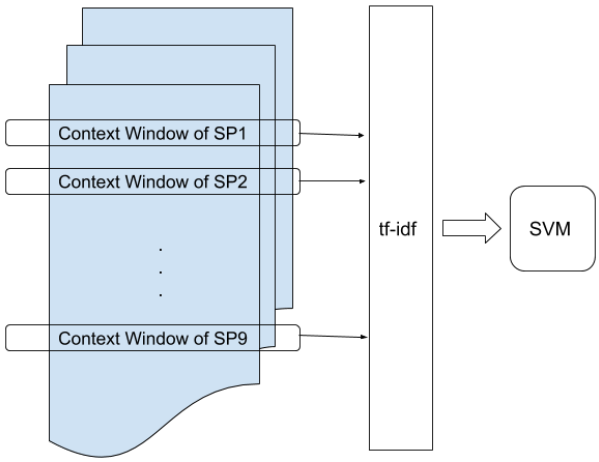

Figure 2: **Narratology-inspired model workflow.**

6 Experimental Setups

Dataset usage. We performed random sampling on both datasets such that 80% is used for training, 10% for validation, and 10% for test.

Baselines. We adopted HAN (Yang et al., 2016), BERT$_{base}$, BERT$_{large}$ (Devlin et al., 2019) as our baselines. Since a script is subdivided into scenes, our HAN implementation, HAN$_{scence}$, uses scence as the second hierarchy instead of sentence.

Method / Dataset	ScriptBase	MSC
HAN$_{scene}$	45.12	45.62
BERT$_{base}$	42.67	46.29
BERT$_{large}$	42.67	46.29
Tfidf-SVM	47.01	59.21
TFIDF-SVM$_{narr}$	**57.43**	**59.21**
+ emo + VAD	56.52	55.29
+ ling + emo + tt	**62.35**	62.73
+ int + ling + emo + clus	60.87	**64.79**

Table 2: **F1 scores (%) of model predictions.**

Implementation details. We use Scikit-learn 0.21.3 to implement feature-based models, and PyTorch 1.3.1 for deep neural models. With Huggingface (Wolf et al., 2019), we overcome BERT's 510-token limit by applying averaging pooling on the sequence of BERT $h_{[CLS]}$ hidden states of subchunks of the script to get a global context vector, and then fine-tune the task end-to-end. And since the binary labels in both datasets are imbalanced, we weight the positive class by inverse frequency of class labels in the training set.

Hyper-parameters. To ensure a fair comparison, we tuned the hyper-parameters for all models. On feature-based models, we performed grid search. For NN models, we use embedding size 100 and Adam optimizer with 0.001 learning rate.

7 Results and Discussion

We report the macro-averaged F1 scores of each model in Table 2, interestingly, from which we see that NN-based document classification methods are no better than our proposed simple narratology-based model. We suppose the length of document could be the main reason, RNNs or transformers may not handle "super long-term depdendencies" well for complex compositions like movie scripts. For NN models, both BERT$_{large}$ and BERT$_{base}$ are better than HAN$_{scene}$, which is expected provided the capacity of BERT is significantly larger than HAN; we are not sure why BERT$_{large}$ did not outperform BERT$_{base}$ by even a slight margin.

In Fig. 3, we show the effect of each individual feature. *Linguistic & Emotional Activity Curve* show improvements on both datasets, and yet the rest do not consistently help, especially on MSC, we think it may be because (1) the tfidf has 500 dimensions so individual feature may be overwhelmed, but, more features combined such

as adding *int+ling+tt* can generate consistent improvements, (2) the efficacy of feature can be dataset-dependent, e.g., we do not observe significant differences in *Arousal* of MSC as in its ScriptBase counterpart (Fig. 1), and so does the classifier. Besides, adding features with negative correlations can damage the performance, e.g., adding *emo & vad*.

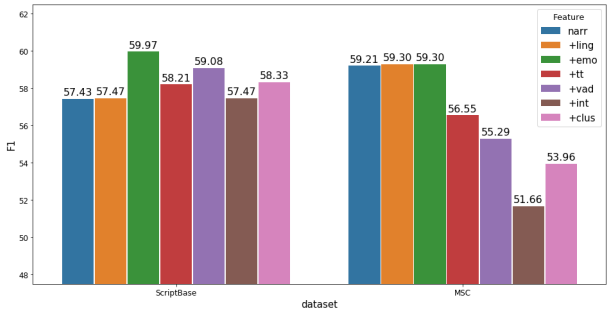

Figure 3: **Individual feature effect.** F1 scores of *Tfidf-SVM*$_{narr}$ and adding proposed features individually.

8 Conclusion and Future Work

We present a novel approach and features to systematically analyze the quality of a screenplay in terms of its festival nomination-worthiness. This can serve as a preliminary tool to help filmmakers in their decision-making, or on the other hand, an objective way for writers to compare their works with others. Our results also show that simple lightweight approach can outperform state-of-the-art document classification methods. This also points out the current deficiency for long document classification research in the community.

In the future, in addition to textual properties, we intend to develop a more fine-grained approach by incorporating more metadata such as gender of characters, film genres, and then experiment on different award categories to evaluate our approach and gain more insights.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Rethinking complex neural network architectures for document classification. In *NAACL-HLT*.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daumé III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *AAAI*, pages 2704–2710.

Prafulla N Dawadi, Diane J Cook, and Maureen Schmitter-Edgecombe. 2016. Modeling patterns of activities using activity curves. *Pervasive and mobile computing*, 28:51–68.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

J. Eliashberg, S. K. Hui, and Z. John Zhang. 2014. Assessing box office performance using movie scripts: A kernel-based approach. *IEEE Transactions on Knowledge and Data Engineering*, 26(11):2639–2648.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Ethan Fast, Bin Bin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *CHI*.

Syd Field. 2007. *Screenplay: The foundations of screenwriting*. Random House LLC.

Stephen Follows, Josh Cockcroft, and Liora Michlin. 2019. Judging screenplays by their coverage: An analysis of 12,000+ unproduced feature film screenplays and the scores they received, revealing what professional script readers think makes a good screenplay.

M. Ghiassi, David Lio, and Brian Moon. 2015. Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications*, 42(6):3176 – 3193.

Philip John Gorinski and Mirella Lapata. 2018. What's this movie about? a joint neural network architecture for movie content analysis. In *NAACL-HLT*.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California. Association for Computational Linguistics.

Rie Johnson and Tong Zhang. 2015. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112, Denver, Colorado. Association for Computational Linguistics.

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17, Montréal, Canada. Association for Computational Linguistics.

Michael T. Lash, Sunyang Fu, Shiyao Wang, and Kang Zhao. 2015. Early prediction of movie success - what, who, and when. In *SBP*.

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

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *SIGIR*.

Robert McKee. 1997. *Substance, Structure, Style, and the Principles of Screenwriting*. New York: HarperCollins.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium. Association for Computational Linguistics.

Saif M. Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M. Mohammad. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. *ArXiv*, abs/1908.10328.

Robert Plutchik. 1980. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3 – 33. Academic Press.

Anil Ramakrishna, Victor R. Martinez, Nikos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *ACL*.

James Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110 1:145–72.

Dean Keith Simonton. 2009. Cinematic success, aesthetics, and economics: An exploratory recursive model.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. *ArXiv*, abs/1909.01007.

T. Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

K.M. Weiland. 2013. *Structuring Your Novel: Essential Keys for Writing an Outstanding Story*. PenForASword.

K.M. Weiland. 2018. Story structure q&a: 6 outstanding questions about structure.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*.