Issue-Focused Documentaries versus Other Films: Rating and Type Prediction based on User-Authored Reviews

Ming Jiang
GSLIS – The iSchool at Illinois
University of Illinois at Urbana-Champaign
{mjiang17@illinois.edu}

Jana Diesner
GSLIS – The iSchool at Illinois
University of Illinois at Urbana-Champaign
{jdiesner@illinois.edu}

ABSTRACT
User-authored reviews offer a window into micro-level engagement with issue-focused documentary films, which is a critical yet insufficiently understood topic in media impact assessment. Based on our data, features, and supervised learning method, we find that ratings of non-documentary (feature film) reviews can be predicted with higher accuracy (73.67%, F1 score) than ratings of documentary reviews (68.05%). We also constructed a classifier that separates reviews of documentaries from reviews of feature films with an accuracy of 71.32%. However, as our goal with this paper is not to improve the accuracy of predicting the rating and type or genre of film reviews, but to advance our understanding of the perception of documentaries in comparison to feature films, we also identified commonalities and differences between these two types of films as well as between low versus high ratings. We find that in contrast to reviews of feature films, comments on documentaries are shorter but composed of longer sentences, are less emotional, contain less positive and more negative terms, are lexically more concise, and are more focused on verbs than on nouns and adjectives. Compared to low-rated reviews, comments with a high rating are shorter, are more emotional and contain more positive than negative sentiment, and have less question marks and more exclamation points. Overall, this work contributes to advancing our understanding of the impact of different types of information products on individual information consumers.

Keywords
Rating prediction; Type prediction; Documentary films; Social impact

1. INTRODUCTION
The impact of media and information products on individuals and small groups has traditionally been measured by surveying people pre and post media exposure [6; 29]. Such surveys as well as in-depth focus group discussions can lead to a deep understanding of a problem domain and people's perception of it, but this process limits scalability [2; 11]. Additionally, peoples' thoughts about information they have consumed can be measured by examining user-authored comments, which can be published on customer review sites, among other sources [19]. Unlike classic interviews with individual information consumers, online reviews can become part of the public discourse about a film or an issue. Analyzing digital reviews eliminates classic issues with surveys, such as answers biased by social desirability, and enables the consideration of large amounts of data over long periods of time. Data mining and Natural Language Processing (NLP) techniques make this approach scalable [5; 19].

In this paper, we focus on a specific subset of media products, namely documentary films, and analyze them in contrast to other types of films. Researchers and practitioners in the field of media impact assessment have defined various sets of impact goals, which often include an increase in public awareness about an issue (other goals are, for example, changes in consumer attitude and behavior, corporate policy, and political action) [3; 12]. Analyzing film reviews offers one way to study a certain subset of the public opinion about a documentary on the micro (i.e., individual) level. We argue that understanding this type of individual engagement contributes to our knowledge about the social impact of documentaries.

Review analysis is a classic subject of study in social computing. Prior work has resulted in knowledge and models for predicting the rating, helpfulness, and sentiment of reviews (details in the Background section) [21; 31; 32]. This work typically does not differentiate between sub-genres of film, but rather uses randomly drawn samples of films from across genres, which is an appropriate solution for results meant to generalize to all types of film. We build upon and extend this prior research in order to a) develop a better understanding of the perception of issue-focused documentaries as a specific genre, and b) identify differences in the ratings and underlying text characteristics of the reviews of documentaries versus other films. Our solution to this task is explained in the Methods section. The findings are presented in the Results section and interpreted in the Discussion section.

We focus on documentaries that address issues of inequality and social justice. These films are often produced as a vehicle to induce change on the individual, group and, ultimately, societal level [6; 7; 18; 22; 30]. The intent of impact co-exists with classic goals of film production, mainly compelling storytelling and appealing cinematography [33]. As stated by a major funder of documentaries, "these stories inspire imaginations, disrupt stereotypes, and help transform attitudes that perpetuate injustice" [15]. However, our understanding of the (aggregated) perception of documentaries is underdeveloped. The research presented in

HT '16, July 10-13, 2016, Halifax, NS, Canada
© 2016 ACM. ISBN 978-1-4503-4247-6/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2914586.2914638.

this paper aims to help to fill this gap. As documentaries are meant to inspire engagement with the public, we are not concerned with improving the accuracy of predicting ratings of films based on reviews, but rather we aim to improve the understanding of differences in the reflection on documentaries versus other films. In summary, in this paper, we use NLP techniques to find answers to the following questions:

1. How does the predictability of ratings of documentaries compare to the predictability of ratings of other types of films?
2. If differences exist, what features of the content of publicly available, user-contributed reviews of films account for these differences?

2. BACKGROUND

The majority of existing work on review mining can be divided into four research areas: First, predicting review helpfulness, which typically is based on explicit votes made by users [17; 19; 21; 26]. Second, rating prediction, with the goal of anticipating user preferences, and third, sentiment analysis, which aims at identifying the features of a product that users like or don't like [8; 10; 28; 38; 39]. Studies on the latter two topics can be further divided into building a) binary classifiers [8; 10; 38], or b) multi-class predictors [30; 32]. Fourth, review summarization, which is meant to identify the gist of information from a set of reviews [23; 26]. All of the abovementioned techniques take the content of the reviews into account, and at least the first three groups may also leverage additional meta-data, such as time stamps and information about the user and the reviewed product.

Our paper falls into the first category, i.e., rating prediction. Scholars have repeatedly confirmed the following features as being helpful for this task: 1) Bag of words, which represents lexical characteristics, 2) parts of speech (POS) tags, which capture shallow syntactic information, and 3) deep syntax information from parse tree, which are useful e.g., to handle negation detection [22; 35]. Prevalent themes identified via topic modeling have also been used as a feature [27]. Finally, features of reviews of other products about which the reviewer of a target product has also written a comment can be considered, e.g., to calibrate a specific rating within a user's profile of rating patterns [28]. Prior work has also identified suitable prediction algorithms or models. Most scholars either use ranking algorithms or apply classification models. For ranking a set of reviews, regression models have emerged as a proper solution [30]. For classification tasks, SVMs [16], Naïve Bayes [31], and Neural Networks [36] have consistently resulted in comparatively highly accurate performance. Overall, most prior rating prediction studies have achieved accuracy rates (F1) of 40% to 85%; average values seem to be around 75% [1; 9; 24; 25; 31; 32].

For the domain of film, we observe a lack of work on the relationship between genre and rating. In this paper, we focus on measuring different features that are appropriate for predicting ratings of documentary reviews, and contrast our findings to results based on other types of films.

3. DATA

We selected documentaries that address different dimensions of inequality and social justice, e.g., economic, political, and cultural issues. The sample is partially based on films which we have considered in our prior related work on social impact assessment of documentaries [13]. We then searched for feature films that address similar topics, have a similar amount of reviews, and have a similar distribution of rating values. In our sample, non-documentaries tend to receive more reviews than documentaries. To keep the amount of reviews on both types of films similar, we considered a smaller set of non-documentaries. We collected reviews on 20 issue-focused documentaries (N=8,090) and 11 other films (N=8,261) from Amazon.com (with their permission). The other films fall into the following genres (a single film can be in multiple categories): drama (7), comedy (4), romance (2), cartoon (2), and science fiction (1). In this paper, we also refer to the films that are not documentaries as "feature films", and acknowledge that this might be an overly general classification or genre. Table 1 provides a summary of the dataset.

Table 1. Dataset statistics

Type of film	Number of reviews	Rating distribution (average and standard deviation)		
		5&4 star (high)	3 star (medium)	1&2 star (low)
Documentaries	8,090	87.45% ±7.22%	5.30%±3.20%	7.25%± 5.52%
Non-documentaries	8,261	82.55% ±8.64%	7.73%±3.74%	9.73%± 5.46%

4. METHOD

Each review is user-rated on a 5-point scale. We consolidated the ratings as follows: a) high ratings, which include all 5 and 4 star reviews; b) medium ratings (3 stars); and c) low ratings (1 and 2 stars). As medium ratings are neither clearly high nor low, and also form the smallest portion of the sample, we disregarded them for further analysis. This implies that we ultimately construct two binary classifiers (one per type or genre of film) for high versus low ratings.

We removed overly short reviews that mainly consisted of stop words. In order to construct a sample that has a similar number of instances of high and low rated reviews for learning, we used the number of low ratings per type (smaller set) as the upper bound for the number of reviews considered per prediction category. We randomly sampled an equally sized corpus of reviews with a high rating from the same films. This process resulted in a total sample size for learning of N=1,000 for documentaries and 1,668 for non-documentaries.

4.1 Feature Selection

All features are extracted from the content of the reviews. Our feature selection is guided by prior work as well as our close reading of a small sample of the reviews. We consider four types of features: meta-data of the texts, text content, regular expressions, and syntax. We use the Stanford NLP toolkit to parse the data and calculate the values specified below [34].

We hypothesize that the text characteristics of the reviews for different genres might differ for two reasons: First, documentaries are a more specifically defined genre than feature films, which might suggest a tighter distribution of unigrams in documentary reviews. Second, these types of film might evoke different styles of engagement, which might be reflected in peoples' writing.

4.1.1 Text Meta-data

We consider the average length of both reviews and of the sentences per review.

4.1.2 Lexical Features

Four content features are considered: salient terms, informativeness, token level sentiment, and transition words. In

order to identify the salient terms, we select the top 250 unigrams per film according to their TF*IDF scores as calculated in Eq1, and normalize the scores by article length (see Eq2).

$$TF * IDF(w, C) = tf(w, C) \times \log\left(1 + \frac{N}{df(w)}\right) \quad (1)$$

$$TF_{Normalized}(w, d) = \frac{tf(w,d)}{\sum_{w' \in d} tf(w',d)} \quad (2)$$

where C represents the corpus per film, w is any term appearing in C, d is any review in the collection, N is the total number of reviews per film, and df(w) is the number of reviews that contain term w. Terms are not syntactically disambiguated for this step.

We also calculate information entropy as it represents the informativeness of a review (Eq3) [38]. Review entropy is computed based on the amount of information that each w carries, which is determined by the w's normalized weight in the review d and corpus C (See Eq4). For this project, we conducted several experiments with different values of λ, and decided to set λ to 0.3.

$$H(d) = \sum_{w \in d} [-p(w) \log_2 p(w)] \quad (3)$$

$$p(w) = \lambda \times \frac{TF*IDF(w,d)}{\sum_{w' \in d} TF*IDF(w',d)} + (1-\lambda) \times \frac{TF*IDF(w,C)}{\sum_{w' \in C} TF*IDF(w',C)} \quad (4)$$

For sentiment identification and quantification, we follow the example of prior studies that use previously constructed and validated dictionaries for this purpose, and chose to use a widely adopted subjectivity lexicon [39]. Terms were syntactically disambiguated for this step. As part of the sentiment analysis, we account for negations by using rule-based negation detection (see Table 2) that relies on deep parsing (as provided by the Stanford NLP Parser).

Finally, we consider transition words [4]. We calculate a) the number of unique transition words per text, and b) the ratio of logical relationships (see Eq5), including addition, introduction, emphasis, conflict/concession, causal, condition, time, and conclusion.

$$Ratio(LR_i) = \frac{\sum_{tw \in LR_i} tf(tw, d)}{\sum_{j=1}^{8} \sum_{wt' \in LR_j} tf(tw', d)} \quad (5)$$

where LR_i represents the i^{th} logical relationship, and tw is any transition word that belongs to the category of LR_i.

Table 2. Rule-based negation detection

Direct negation rules	Indirect negation rules
neg(VB/JJ, not)	neg(w, not) + amod(w, JJ) => not JJ
	neg(w, not) + xcomp(JJ, w) => not JJ
	neg(w, not) + admod(RB, w) => not RB

4.1.3 Regular Expression

We calculate the ratio of question marks and exclamation points per review (also by using the Stanford NLP Parser).

4.1.4 Syntax Features

We identify the POS per word and parse tree constituents per sentence. For each review, we calculate a) the number of unique POS tags, and b) the ratio of nouns (i.e., NN, NNS, NNP & NNPS), verbs (i.e., VB, VBD, VBG, VBN, VBP & VBZ), adjectives (i.e., JJ, JJR & JJS), and adverbs (i.e., RB, RBR & RBS).

4.2 Learning and Evaluation

Following the example of prior studies, we use a SVM with a radial kernel for learning. The classifier was implemented using the R package e1071 [14]. For our experiments, we conducted 10-fold cross validations, and report the averaged results. For assessing prediction accuracy, we use the standard metrics of precision, recall, and the F1 score.

5. RESULTS

5.1 Classification Performance

Based on our data, features, and learning method, we find that ratings of non-documentaries can be predicted with higher accuracy (73.67%, F1 score) than ratings of documentaries (68.05%) (Table 3). This difference could be due to differences in the sample size per type of film (larger for non-documentaries), or could mean that high versus low-rated reviews are more distinct for feature films than for documentaries. Our accuracy rates for rating prediction of feature films are a little lower than in prior work (about 75%, see Background section for details), while no point of comparison exists specifically for documentaries.

Especially for feature films, precision is higher than recall when

Table 3. Accuracy of rating prediction per type of film (average and standard deviation)*

Used Features	Recall		Precision		F1	
	Docu	Non-Docu	Docu	Non-Docu	Docu	Non-Docu
Review length (RL)	78.19% ±0.048	68.15% ±0.059	59.48% ±0.051	66.02% ±0.042	67.43%±0.040	66.87%%±0.034
Avg. sentence length	48.25% ±0.077	49.76% ±0.063	61.43% ±0.081	65.31% ±0.060	53.64%±0.066	56.32%%±0.054
Unigram	70.29% ±0.091	63.78% ±0.042	61.54% ±0.060	65.96% ±0.067	65.13%±0.051	64.64%%±0.038
Entropy	70.13% ±0.085	63.31% ±0.051	61.70% ±0.075	65.46% ±0.056	65.52%±0.073	64.30%%±0.048
Sentiment%	69.42% ±0.072	66.27% ±0.067	63.41% ±0.054	76.48% ±0.033	65.98%±0.043	70.95%%±0.053
Transition words	60.99% ±0.091	54.39% ±0.088	61.76% ±0.062	67.67% ±0.054	61.00%±0.061	60.01%%±0.068
Question mark% (Q)	**91.99% ±0.041**	**93.47% ±0.031**	55.86% ±0.049	55.44% ±0.032	69.44%±0.046	69.54%%±0.029
Exclamation mark%	19.92% ±0.032	13.89% ±0.023	**74.93% ±0.119**	78.64% ±0.098	31.25%±0.043	23.58%%±0.036
POS	61.52% ±0.055	60.52% ±0.025	64.51% ±0.073	70.46% ±0.059	62.78%±0.051	64.97%%±0.026
Num_Sentiment+Negative% (Senti_N)	74.11% ±0.056	59.36% ±0.045	65.93% ±0.077	78.72% ±0.048	69.41%±0.044	67.55%%±0.035
Num_Sentiment +Positive% (Senti_P)	77.67% ±0.063	70.57% ±0.050	60.83% ±0.079	70.39% ±0.056	67.91%±0.056	70.18%%±0.022
Senti_N+Q+RL	75.05% ±0.084	60.10% ±0.063	66.12%±0.089	**79.05% ±0.059**	**69.91%±0.070**	68.08%%±0.050
Senti_P+Q +RL	78.99% ±0.061	72.65% ±0.053	62.26%±0.060	71.87% ±0.042	69.39%±0.042	72.04%%±0.024
All	**68.16% ±0.100**	**71.84% ±0.050**	**68.82% ±0.055**	**75.86% ±0.048**	**68.05%±0.059**	**73.67%%±0.037**

* highest value per column marked in bold

Table 4. Accuracy of type of film prediction (average and standard deviation)*

Recall	Precision	F1
72.19% ±0.013	70.50% ±0.019	71.32% ±0.013

* using all texts and features

using all features (Table 3). Our model is more likely to predict a truly high rating as a low rating than vice versa, which means low ratings are easier to recognize, while high ratings are more ambiguous.

The ratio of question marks is the strongest individual feature with respect to recall for both types of film (91.99% for documentary reviews, 93.47% for feature films reviews). For precision, the strongest feature for documentary reviews is the ratio of exclamation points (74.93%), and for feature films, it is a combination of the amount of words with a negative sentiment, the ratio of question marks, and review length (79.05%).

For combining multiple features for learning, the F1 values suggest that the number of negative sentiment terms plus the ratio of question marks plus review length is the best feature set for classifying documentary reviews, while for non-documentary reviews, combining all features results in the highest accuracy rates. We also find negative sentiment to be more indicative of documentary reviews, and positive sentiment to be a better predictor for non-documentary reviews. At least two explanations seem plausible, but require further testing for confirmation. First, documentary reviews might be written by a more critical audience. Second, and independent of the reviewers' perception and style, documentary reviews might address or represent the severity of given social justice issues.

Instead of building two binary classifiers, the prediction task solved in this paper can also be approached as a 4-label classification problem (high versus low-rated reviews of documentaries versus other films). Using the same sample of documentaries as for the previous task, and an equally sized sample of non-documentary reviews, we tested this approach by using all introduced features, and achieved an overall F1 score of 44.55%, which is considerably lower than the accuracy obtained with the prior approach.

Finally, we trained a binary classifier that aims to tell apart reviews of documentaries versus other films (Table 4). Using the full set of reviews and features, we obtained an accuracy rate of 71.32% (F1) for distinguishing reviews per type, regardless of the rating. This finding suggests that reviews per genre have distinct characteristics, which are analyzed in more detail in the next section.

5.2 Feature Analysis

In this section, we analyze the differences between the set of documentaries reviews versus feature film reviews based on the entire corpus.

5.2.1 Feature Comparison by Rating

Most of the findings in this section are shown in Table 5. In our sample, high-rated reviews are considerably shorter than low-rated reviews. This might indicate that agreement or excitement get expressed with brevity, while disagreement or disappointment are associated with more detailed explanations.

Low-rated reviews contain more question marks and less exclamation points than high-rated reviews. This suggests that people raise more questions in critical or negative reviews, and emphasize their opinion more in positive reviews.

Table 5. Feature comparison by rating group

Selected Feature	4&5 stars (high)		1&2 stars (low)	
	Docu	Non-Docu	Docu	Non-Docu
Review length	76	78	145	140
Sentiment%	12.82%	13.88%	10.57%	10.22%
Positive %	41.15%	56.84%	30.02%	33.31%
Negative %	22.11%	16.02%	37.76%	37.31%
Conflict%	8.13%	9.51%	12.00%	15.13%
Emphasis%	12.88%	10.05%	8.80%	10.17%
Question mark%	1.65%	1.38%	5.68%	4.95%
Exclamation mark%	13.59%	14.52%	5.08%	6.14%

As one might expect, higher ratings correlate with more positive and less negative sentiment, and also with higher emotionality. The gap between the ratio of positive to negative words decreases with decreasing ratings.

5.2.2 Feature Comparison by Type of Film

Most of the findings in this section are represented in Table 6. On average, non-documentary reviews are longer, but composed of shorter sentences than documentary reviews. This could indicate that feature film reviews are easier to write. Also, documentary reviews have a slightly higher entropy than feature film reviews.

Across types of film, in total, reviews contain more positive than negative terms (Table 6). This finding suggests a general level of courtesy and politeness among laymen film reviewers. Compared to documentaries, comments on non-documentaries have a slightly higher ratio of sentiment words, a considerably larger ratio of positive terms, and a lower portion of negative terms. These findings indicate that reviews of feature films are more emotional and enthusiastic. A possible explanation for this effect

Table 6. Feature comparison by type of film

Feature Type	Feature	Docu	Non-Docu
Text meta-data	Review length (in words)	82.6310	91.4873
	Avg. sentence length	14.0075	12.6440
Regular expressions	Question mark%	1.98%	1.90%
	Exclamation mark%	12.53%	12.75%
Sentiment	Sentiwords%	12.60%	13.23%
	Positive%	40.18%	53.33%
	Negative%	23.45%	19.23%
POS	# Unique POS tags	14.6671	13.7680
	NN%	24.38%	26.42%
	VB%	18.42%	15.58%
	JJ%	11.20%	14.11%
	RB%	6.24%	6.22%
Transition words	Unique transition words%	6.85%	6.94%
	Addition%	43.79%	40.02%
	Introduction%	0.77%	0.60%
	Emphasis%	12.49%	9.80%
	Conflict%	8.93%	11.01%
	Causal%	5.64%	6.07%
	Condition%	4.11%	3.22%
	Time%	5.40%	5.12%
	Conclusion%	0.18%	0.20%
Unigram	Top unigram 1	0.0112	0.0108
	Top unigram 2	0.0179	0.0100
	Top unigram 3	0.0173	0.0181
	Top unigram 4	0.0212	0.0141
	Top unigram 5	0.0118	0.0119
Information quantity	Entropy	1.7525	1.7387

Table 7. Example for top 20 unigrams*

Fed Up (Docu)	sugar, food, *movie, documentary*, people, *film*, *watch*, fat, health, *great*, industry, foods, eat, *good*, government, eye, eating, *informative*, obesity, children
War Horse (Non-Docu)	horse, *film*, war, *joey*, *spielberg*, *story*, horses, *great*, *good*, albert, love, *movies*, *scenes*, *watch*, *time*, family, *loved*, boy, animal, man

* words not central to key topic of film in italics

might be the way in which similar topics are presented in documentaries versus feature films.

With respect to syntax, people use comparatively more nouns and adjectives in non-documentary reviews, and more verbs in documentary reviews. This might suggest that non-documentary reviews are more about objects or social entities and their modifiers (e.g., great film!), while documentary reviews might focus more on activities. More analyses are needed to test this assumption, but the latter finding is a desirable outcome for impact creators.

Finally, the unigram analysis shows a stronger focus tendency in documentary reviews. This finding might reflect the fact that the themes addressed in documentaries and/or their reviews are more focused on specific topic, while individual feature films might cover a broader scope of topics. To illustrate this point, we provide an example: We show the top 20 unigrams (based on TF*IDF) from two randomly selected films in Table 7. This comparison reveals that more than half of the unigrams occurring in reviews on *Fed Up* (a documentary) focus on junk food and associated health issues (about 13 of the terms), which is the main issue of the film. For *War Horse* (a non-documentary), the top unigrams represent several topics, including the actual theme of the film (about 8 of the 20 terms, including "horse", "war", "horses"), the leading actor ("joey"), the director ("spielberg"), and other themes.

6. CONCLUSION AND DISCUSSION

We have built two binary classifiers that predict high versus low ratings of reviews of issue-focused documentaries versus other films with 68.05% and 73.67% accuracy (F1), respectively. We also constructed a classifier that separates reviews of documentaries from reviews of feature films with an accuracy of 71.32%. However, as our goal with this paper is not to improve rating and type prediction accuracy, but to advance our understanding of the perception of documentaries in comparison to feature films, we also identified commonalities and differences between these two genres as well as between low versus high ratings in general: In contrast to reviews of feature films, comments on documentaries are shorter but composed of longer sentences, are less emotional, contain less positive and more negative terms, are lexically more concise, and are more focused on verbs than nouns and adjectives. Compared to low-rated reviews, comments with a high rating are shorter, are more emotional and contain more positive than negative sentiment, and have less question marks and more exclamation points.

Our work has several limitations. First, by aggregating all reviews per film, we gain only a general sense of the users' opinions and engagement with a film. Time slicing the reviews would allow for analyzing changes in the perception of a film over time. Second, we apply a very coarse definition of feature films. Our work could be improved by further splitting feature films into more precise genres. Third, most reviews on Amazon are authored by laymen. We have started to complement this work by also considering reviews from expert critics, which are typically published in traditional print media.

ACKNOWLEDGMENTS

This work is supported by the FORD Foundation, JustFilms division, grant 0155-0370. We thank Amazon.com for giving us permission to collect and use customer review data from their site. We thank Chieh-Li Chin and Rezvaneh Rezapour from the iSchool at UIUC for their help with and advice about this paper.

7. REFERENCES

[1] Adomavicius, G. and Kwon, Y., 2007. New recommendation techniques for multicriteria rating systems. *Intelligent Systems, IEEE 22*, 3, 48-55.

[2] Bernard, H.R., 2012. *Social Research Methods: Qualitative and Quantitative Approaches*. Sage.

[3] Britdoc, *The end of the line. A social impact evaluation.* http://animatingdemocracy.org/resource/end-line-social-impact-evaluation.

[4] Campbell, G.M., Buckhoff, M., and Dowell, J.A., Transition Words. https://msu.edu/~jdowell/135/transw.html.

[5] Chaovalit, P. and Zhou, L., 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences, (HICSS'05)*. IEEE, 112c.

[6] Chattoo, C.B., 2014. *Assessing the Social Impact of Issue-Focused Documentaries: Research Methods and Future Considerations.* Center for Media and Social Impact, School of Communication at American University.

[7] Clark, J. and Abrash, B., 2011. *Social Justice Documentary: Designing for Impact.* Center for Social Media, School of Communication at American University http://www.centerforsocialmedia.org/designing-impact.

[8] Cui, H., Mittal, V., and Datar, M., 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the 21st International Conference on Artificial intelligence, (AAAI'06)*. 1265-1270.

[9] De Albornoz, J.C., Plaza, L., Gervás, P., and Díaz, A., 2011. A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating. In *Advances in Information Retrieval.* Springer, Berlin Heidelberg, 55-66.

[10] Devitt, A. and Ahmad, K., 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, (ACL'07).* Association of Computational Linguistics, 25-27.

[11] Diesner, J., Kim, J., and Pak, S., 2014. Computational impact assessment of social justice documentaries. *Metrics for Measuring Publishing Value: Alternative and Otherwise 17*, 3.

[12] Diesner, J. and Rezapour, R., 2015. Social Computing for Impact Assessment of Social Change Projects. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, (SBP'15).* Springer, 34-43.

[13] Diesner, J., Rezapour, R., and Jiang, M., 2016. Assessing public awareness of social justice documentary films based

on news coverage versus social media. In *Proceedings of the iConference*.

[14] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A., 2011. Misc functions of the department of statistics (e1071), TU Wien. In *R package 1*, Version: 1-6.

[15] Ford Foundation, Just Films. http://www.fordfoundation.org/work/our-grants/justfilms.

[16] Ganu, G., Elhadad, N., and Marian, A., 2009. Beyond the stars: improving rating predictions using review text content. In *Proceedings of the 12th International Workshop on the Web and Databases, (WebDB'09)*. 1-6.

[17] Green, D. and Patel, M., 2013. *Deepening Engagement for Lasting Impact: A Framework for Masuring Media Performance and Results*. John S. and James L. Knight Foundation and Bill & Melinda Gates Foundation.

[18] Hong, Y., Lu, J., Yao, J., Zhu, Q., and Zhou, G., 2012. What reviews are satisfactory: novel features for automatic helpfulness voting. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval, (SIGIR'12)*. ACM, 495-504.

[19] Hu, M. and Liu, B., 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining, (KDD'04)*. ACM, 168-177.

[20] Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M., 2006. Automatically assessing review helpfulness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP'06)*. Association for Computational Linguistics, 423-430.

[21] Knight Foundation, 2011. *Impact: A Guide to Evaluating Community Information Projects*. http://www.knightfoundation.org/publications/impact-practical-guide-evaluating-community-inform.

[22] Li, S., Zhang, H., Xu, W., Chen, G., and Guo, J., 2010. Exploiting combined multi-level model for document sentiment analysis. In *Proceedings of the 20th International Conference on Pattern Recognition, (ICPR'10)*. IEEE, 4141-4144.

[23] Liu, C.-L., Hsaio, W.-H., Lee, C.-H., Lu, G.-C., and Jou, E., 2012. Movie rating and review summarization in mobile environment. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 42*, 3, 397-407.

[24] Liu, J. and Seneff, S., 2009. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP'09)*. Association for Computational Linguistics, 161-169.

[25] Liu, Y., Huang, X., An, A., and Yu, X., 2008. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the 8th IEEE International Conference on Data Mining, (ICDM'08)*. 443-452.

[26] Ly, D.K., Sugiyama, K., Lin, Z., and Kan, M.-Y., 2011. Product review summarization from a deeper perspective. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, (JCDL'11)*. ACM, 311-314.

[27] Mcauley, J. and Leskovec, J., 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems, (RecSys'13)*. ACM, 165-172.

[28] Mukherjee, S., Basu, G., and Joshi, S., 2013. Incorporating author preference in sentiment rating prediction of reviews. In *Proceedings of the 22nd International Conference on World Wide Web, (WWW'13)*. ACM, 47-48.

[29] Napoli, P., 2014. *Measuring Media Impact: An Overview of the Field*. Media Impact Project, USC Annenberg Norman Lear Center.

[30] Pang, B. and Lee, L., 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, (ACL'05)*. Association for Computational Linguistics, 115-124.

[31] Pang, B., Lee, L., and Vaithyanathan, S., 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing-Volume 10, (EMNLP'02)*. Association for Computational Linguistics, 79-86.

[32] Qu, L., Ifrim, G., and Weikum, G., 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics, (COLING'10)*. Association for Computational Linguistics, 913-921.

[33] Rose, F., 2012. *The Art of Immersion: How the Digital Generation Is Remaking Hollywood, Madison Avenue, and the Way We Tell Stories*. W.W. Norton & Company, New York, NY.

[34] Socher, R., Bauer, J., Manning, C.D., and Ng, A.Y., 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, (ACL'13)*. Association for Computational Linguistics, 455-465.

[35] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M., 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics. 37*, 2, 267-307.

[36] Tang, D., Qin, B., Liu, T., and Yang, Y., 2015. User modeling with neural network for review rating prediction. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence, (IJCAT'15)*. 1340-1346.

[37] Turney, P.D., 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, (ACL'02)*. Association for Computational Linguistics, 417-417.

[38] Weaver, W. and Shannon, C.E., 1949. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois.

[39] Wilson, T., Wiebe, J., and Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, (HLT'05)*. Association for Computational Linguistics, 347-354.