Says Who…? Identification of Expert versus Layman Critics' Reviews of Documentary Films

Ming Jiang
School of Information Sciences
University of Illinois at Urbana-Champaign
{mjiang17@illinois.edu}

Jana Diesner
School of Information Sciences
University of Illinois at Urbana-Champaign
{jdiesner@illinois.edu}

Abstract

We extend classic review mining work by building a binary classifier that predicts whether a review of a documentary film was written by an expert or a layman with 90.70% accuracy (F1 score), and compare the characteristics of the predicted classes. A variety of standard lexical and syntactic features was used for this supervised learning task. Our results suggest that experts write comparatively lengthier and more detailed reviews that feature more complex grammar and a higher diversity in their vocabulary. Layman reviews are more subjective and contextualized in peoples' everyday lives. Our error analysis shows that laymen are about twice as likely to be mistaken as experts than vice versa. We argue that the type of author might be a useful new feature for improving the accuracy of predicting the rating, helpfulness and authenticity of reviews. Finally, the outcomes of this work might help researchers and practitioners in the field of impact assessment to gain a more fine-grained understanding of the perception of different types of media consumers and reviewers of a topic, genre or information product.

1 Introduction

Product reviews help customers to make purchase decisions, and producers to improve and develop goods (Hu & Liu, 2004; Kim et al., 2006; Mudambi & Schuff, 2010). Scalable NLP-based solutions have been developed to support various aspects of these decision-making processes:

(1) Describing, understanding and anticipating product ratings can help manufacturers to comprehend a market. Ranking reviews and reviewers further aids this step. The rating values per review are typically user-generated, ordinal variables, often on a 5-point scale (Jiang & Diesner, 2016; Pang & Lee, 2005).

(2) Identifying the trustworthiness or authenticity of reviews can assist in separating authentic from fudged reviews (Jindal & Liu, 2007; Jindal et al., 2010; Wu et al., 2010). Predicting this feature is more challenging than the previously mentioned ones as authenticity values are not explicitly provided by reviewers or readers, but need to be inferred from the content of reviews and related metadata.

(3) Predicting whether a review was written by an expert or a layman helps to differentiate the impact of the electronic word-of-mouth on the e-marketplace. McAuley and Leskovec (2013) studied the change of reviewers' expertise over time in order to improve personal recommender systems. Knowing the type of reviewer can also assist with asserting the credibility of reviewers (Basuroy et al., 2003; Flanagin & Metzger, 2013; Liu et al., 2008).

Besides commercially motivated analyses of product reviews, assessing the impact of information products such as books, films and other works of art on individuals, groups and society is another domain where knowing the type of author can be beneficial. In the context of impact assessment, *expert critics* (short experts in this paper) can be conceptualized as people with high standards of integrity and an extrinsic motivation for this task, such as writing reviews as part of their jobs as journalists. *Laymen reviewers* can be considered as ordinary customer who are intrinsically motivated to voluntarily provide this type of user-generated content based on their personal experience and points of view (Amblee & Bui, 2007; Chattoo & Das, 2014; Napoli, 2014; Rezapour & Diesner, 2017). We acknowledge the possibility that laymen might write expert-level reviews and vice versa.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2122–2132, Osaka, Japan, December 11-17 2016.

Creators and funders of works of art can use the knowledge about the type of writer to evaluate the impact of information products on the public, e.g. in terms of knowledge diffusion, framing, and sentiment. To provide some better understanding of this process, in this paper, we develop a binary classifier that predicts whether a review of a documentary film was authored by an expert or a layman. We focus on the domain of issue-focused documentaries to complement work based on feature films and box-office blockbusters. Our work also complements prior knowledge gained from studies that predict reviewer expertise based on personal ratings (Amblee & Bui, 2007; Flanagin & Metzger, 2013; Plucker et al., 2009), and/or the online behaviour of reviewers (Liu et al., 2008). We hypothesize that the type of reviewer can be inferred from characteristics of the text data, and address the following research question: Do different text patterns exist in reviews authored by experts versus laymen? If so, what features are unique to each group?

To the best of our knowledge, our study is the first one to apply machine-learning methods to computationally detect the type of author based on the content of reviews. We achieve an overall prediction accuracy of about 90.70% (F-measure), and explain the characteristics of each predicted class (expert versus layman).

The remainder of this paper is organized as follows. We review related work in section 2. Our corpus is described in section 3, and the methods in section 4. In the results section (5), we identify characteristics of each type of author and provide an error analysis. Finally, conclusions and future work are discussed in section 6.

2 Related Work

Nelson (1970, 1974) divides products into "search goods," i.e., tangible objects like cars, and "experience goods," i.e., intangible objects like films. While it might be possible to objectively evaluate search goods, e.g. in terms of their form, function and behaviour, rating experience goods, which are the subject of this study, may involve more personal perspectives, opinions, emotions and subjective judgment (Liu et al., 2008).

The majority of prior NLP-based solutions to commercially inspired review mining tasks can be divided into three groups: First, studies that predict the rating of products and the helpfulness of reviews (Ghose & Ipeirotis, 2011; Kim et al., 2006; Liu et al., 2008; Mudambi & Schuff, 2010; Yang et al., 2015; Zhang et al., 2015). These two tasks are fairly straightforward because user-generated ground-truth data on these features is available. Second, work that identifies the sentiment or opinion entailed in reviews. Knowledge about this feature might also help to explain or predict the former two features. This task requires the labelling of reviews with (values for) sentiment or opinion categories. Building such predictors is typically approached by using deterministic (look-up dictionaries) and/or probabilistic NLP techniques (de Albornoz et al., 2011; Pang & Lee, 2005; Turney, 2002). Third, work that focuses on summarizing the content or the gist of reviews to reduce the complexity of large text corpora (Hu & Liu, 2004; Li et al., 2010; Zhuang et al., 2006).

Studying reviewer expertise, which is the focus of this paper, is a minor branch in current review mining research. In prior work, this problem has mainly been approached by a) using empirical statistical investigations, such as counting average rating scores of experts versus novices, or correlating rating values with product consumption, b) conducting content analysis of reviews (de Jong & Burgers, 2013; Mackiewicz, 2009), and c) computational identifying the level of reviewers' expertise. As an example for the last type, Liu and colleagues (2008) used "reviewer expertise" as one of three variables for identifying the helpfulness of movie reviews. The authors operationalized expertise as frequent and highly positive-rated reviews per author and per pre-defined film genre. Their other two features were writing style and review timeliness. Combining all three features in a non-linear regression resulted in a helpfulness prediction accuracy of 71.2% (F-measure). The isolated contribution of the expertise feature was 51.8% (F-measure). In another study, which also falls into the last category, McAuley and Leskovec (2013) showed that users become more experienced in developing their taste for experience goods (tested for the product categories of beer, wine, fine foods and movies) with over-time exposure to these products. The authors found that the accuracy for predicting item ratings increases when users had higher levels of experience or expertise.

Our work differs from prior studies in that we focus on predicting reviewer expertise as a binary variable based on text-based features of reviews of issue-focused documentaries. The primary goal of

Abbreviation	Documentary	#Expert Reviews	#Valid Expert Reviews	#Layman Reviews	#Total Valid Reviews
SPSZM	Super Size Me	770	166	727	893
INJO	Inside Job	246	68	905	973
FOIN	Food Inc	129	65	2707	2772
GTKER	The Gatekeepers	85	47	178	225
TCOV	The Cove (fishing film)	78	45	485	530
CTFR	Citizenfour	97	44	238	282
AOKI	The Act of Killing	130	39	100	139
BLKFSH	Blackfish	69	35	1171	1206
EOTL	The End of the Line	40	33	67	100
FBCR	5 Broken Cameras	40	28	119	147
TTDS	Taxi To the Dark Side	220	27	52	79
HILI	House I Live In	45	26	221	247
HTSAP	How to Survive a Plague	42	19	79	98
HABA	Hell and Back Again	30	19	67	86
DWAR	Dirty Wars: The World Is a Battlefield	45	17	416	433
IVWAR	The Invisible War	36	16	231	247
PL3P	Paradise Lost 3 Purgatory	17	10	125	135
PAPR	Pandora's Promise	15	10	41	51
PDBTH	Pray the Devil Back to Hell	12	5	51	56
TALD	Through a Lens Darkly	10	4	10	14
SUM		**2156**	**723**	**7990**	**8713**

Table 1: Corpus statistics

this study is to detect indicative text features that can differentiate reviews written by experts from laymen.

3 Data

We collected a dataset that contained ground-truth or gold-standard data, i.e., expert versus layman reviews, for 20 documentaries. The films were selected based on their coverage of main social justice issues (as defined by philanthropic funders), including environmental issues, politics, public health, gender and ethnicity (Diesner et al., 2016). Table 1 shows the list of selected films, their abbreviation used in this paper, and the number of reviews per category.

Based on our reading of reviews on several popular film-rating sites, such as Rotten Tomatoes, Metacritic, Amazon and YouTube, we assume that layman reviews are mainly provided voluntarily. For these reviews, the full texts are provided on these websites. However, for expert reviews, only snippets and a link to the original source (e.g., major newspapers) are typically displayed. Due to copyright regulations and the terms of service for these pages, we could not access expert reviews from these review sites. Alternatively, we used LexisNexis Academic to collect comments written by professional critics that were published in newspapers and other sources. For these searches, the queries contained the film's title, name(s) of the director and/ or producer, and the keyword "review". The latter two items mainly served as disambiguators. For laymen reviews, we collected customer reviews from Amazon after obtaining Amazon's permission for this procedure. Even though reviews per author type were collected from a different platform (customer reviews from Amazon, expert reviews from LexisNexis Academic), we argue that the source does not determine or predict the type of author for the following reason: Many of these platforms list both types of reviews side by side, on the same platform. In other words, expert reviews from sources like Rotten Tomatoes are not written by Rotten Tomatoes, but come from the same sources that we used for our study – e.g., major newspapers.

The data collection involved some challenges. First, manual inspection of each article from LexisNexis was unavoidable as many texts were (soft) duplicates or poor fits, e.g., comments on multiple films with the target film being only briefly mentioned. We manually eliminated duplicates, false positives and poor fits. In the end, 33.53% of the downloaded reviews were judged as valid data points for this study. The resulting number of valid instances per class is also shown in Table 1.

Second, the number of laymen reviews exceeds that of expert reviews. Therefore, for learning, we used the smaller set (i.e., expert reviews) as the defining upper bound for the number of instances considered per class, and randomly sampled an equally-sized number of reviews from the larger set.

4 Method

4.1 Features

Our features selection is guided by prior work that have shown that different aspects of writing style are useful indicators of review helpfulness and reviewer expertise. Based on this prior work, we chose three types of features (discussed in detail below): length features, lexical features and syntactic features.

The content of all considered reviews (N=1446; 723 per class) was pre-processed via stop word removal, stemming, and converting capitalization to lower case[1] for most lexical features except for sentiment analysis. We tested the impact of each routine on feature construction and prediction performance (F-measure), and selected the abovementioned techniques as they contributed most strongly to prediction performance.

4.1.1 Length Features

The length of both reviews and sentences per reviews were considered (Review length, Average sentence length) and computed by using the Stanford Parser (De Marneffe et al., 2006).

4.1.2 Lexical Features

As word choice may also characterize or correlate with each type of reviewer, we leveraged the top 250 unigrams according to the TF*IDF metric (unigram) as shown in Equation 1.

$$weight(w, C_f) = tf(w, C_f) \times idf(w) = c(w, C_f) \times \log\left(1 + \frac{N}{df(w)}\right) \quad (1)$$

$$weight(w, d) = tf(w, d) \times idf(w) = c(w, d) \times idf(w) \quad (2)$$

In this equation, C_f represents the corpus of all reviews per film, w is any term in C_f, c(w, C_f) is the number of occurences of w in C_f, N is the total number of reviews in the collection of a film, and *df(w)* is the number of reviews within the corresponded collection in which w appears.

Equation 2 calculates the TF*IDF per unigram per review, where d is the content per review.

We also used the informativeness per review as a feature (Equation 4) (Weaver & Shannon, 1949) by calculating information entropy. This metric is based on the average amount of information that each w carries per review as well as in the whole corpus per film, respectively (see Equation 3). The calculation is determined by the w's normalized weight in review d and corpus C. As we focus more on the amount of information carried by w in corpus, the ratio parameter λ was set to 0.3 after experimenting with various values.

$$p(w) = \lambda \times \frac{weight(w, d)}{\sum_{w' \in d} weight(w', d)} + (1 - \lambda) \times \frac{weight(w, C_f)}{\sum_{w' \in C} weight(w', C_f)} \quad (3)$$

$$H(d) = \sum_{w \in d} [-p(w) \log_2 p(w)] \quad (4)$$

In addition, the emotionality of reviews has been shown to correlate with formal (more neutral) versus informal (more emotional) writing styles (Hu & Liu, 2004; Jiang & Diesner, 2016; Kim et al., 2006). To calculate emotionality, we reused the previously built, evaluated and widely used MPQA Subjectivity Lexicon (Wilson et al., 2005). Using this external lexical resources, we identified sentiment-loaded terms, summed them up per text, and normalized the sum of the number of sentiment words per valence type by text length (Sentiment%).

[1] Implemented by using an open-source package:
https://github.com/ijab/trec_file_ir/tree/master/bin/edu/pitt/sis/infsci2140/analysis

TR	Definition	Examples
Addition	Provide similar or further information	and, also, or, further
Introduction	Illustrate an argument with a detailed instance	for example, such as
Emphasis	Underline an argument	Even, very especially
Concession	Counter a previous argument	but, however, although
Causality	Describe cause and effect	because, since
Condition	Explain a precondition	if, unless
Order	Sequentially order	before, after
Summary	Conclusion	in a word

Table 2: Selected transition relationships

Syntax	Usage
Aux	Identify clause with non-main verb
Auxpass	Identify passive voice of clause with non-main verb
Csubj	Identify subject clause
Csubjpass	Identify passive voice of subject clause
Dobj	Identify direct object of clause
Iobj	Identify indirect object of clause
Nsubj	Identify nominal subject in clause
Nsubjpass	Identify passive nominal subject in passive clause
Mark	Identify finite clause that subordinate to another clause

Table 3: Selected syntax dependencies

Finally, we considered transition words and phrases to capture text cohesion, i.e., how ideas within a review relate to each other. We also leveraged an external lexical resource for this feature (Campbell et al.). Table 2 shows the list of main transition relationships (TR) used in this paper. We counted the number of transitions per review and normalized the value by review length (Transition%). We also calculated the ratio of each type of TR to capture individual preferences among TR per text (see Equation 5), where t is any transition term that appears in the review d and belongs to the i^{th} type of TR. N_{t_d} notes total number of transition terms in d.

$$Ratio(TR_i, d) = \frac{\sum_{t \in TR_i} tf(t,d)}{N_{t\_d}} \qquad (5)$$

4.1.3 Syntax Features

The Stanford POS tagger was used to assign a single best fitting grammatical function (part of speech or POS) to every token. Per review, we calculated: 1) POS diversity, i.e., the number of unique POS tags, and normalized the value by total 36, which is the total number of POS tags considered by the tagger (Marcus et al., 1993), and 2) the prevalence of content bearing terms (see Equation 6), where tag is any POS tag that belongs to the i^{th} type of content words $Content\_W$ in review d. N_{tag_d} represents the total number of POS tags appeared in d. For each type of content words, we considered a set of POS tags shown as below:

- Nouns (i.e., NN, NNS, NNP & NNPS)
- Verbs (i.e., VB, VBD, VBG, VBN, VBP & VBZ)
- Adjectives (i.e., JJ, JJR & JJS)
- Adverbs (i.e., RB, RBR & RBS)

$$Ratio(Content\_W_i, d) = \frac{\sum_{tag \in Content\_W_i} count(tag,d)}{N_{tag\_d}} \qquad (6)$$

Beyond the token level syntax features, we further used the Stanford NLP Parser to take the grammatical functions of words on the sentence level into account. Similar to our approach for using POS tags, for each review, we calculated: 1) syntax label diversity, where we normalized the number of unique syntax dependencies which appear in each review by the total number of dependency relations (N=48) given in the Stanford parser (De Marneffe et al., 2006), and 2) the ratio of each selected syntax dependency (see Table 3) to all dependencies occurring per review; using the Stanford typed dependencies for this task (De Marneffe & Manning, 2008).

4.2 Learning and Evaluation

After experimenting with various learning algorithms and observing SVM outperforming Naïve Bayes, we decided to present results based on training an SVM with a radial kernel. The classifier was implemented using the R package e1071 (Dimitriadou et al., 2011).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \begin{cases} Precision = \frac{TP}{TP+FP} \\ Recall = \frac{TP}{TP+FN} \end{cases} \tag{7}$$

In order to assess the performance of our features, we conducted a 10-fold cross validation, where we used all reviews (expert and non-expert) for 18 films for training, and documents from the remaining 2 films for testing. To create comparatively similarly sized folds, we sorted films by decreasing numbers of reviews, and iteratively combined the two films from each end into one fold. This non-standard way of partitioning the data was chosen to enable result interpretation and error analysis not only on the class label basis, but to also be able to see if certain films, e.g. on certain topics or from different release dates, impact predictability.

We evaluate the performance of the proposed approach using the standard metrics of precision, recall and the F1 score (see Equation 7). Since we code the class labels with 0 for experts and 1 for laymen, TP (i.e., true positives) represents the number of layman reviews that are correctly predicted while FP (i.e., false positives) is the number of reviews that are mistakenly predicted as layman reviews. TN (i.e., true negatives) and FN (i.e., false negatives) are defined in the same way, but for expert reviews.

5 Results and Analysis

5.1 Experimental Results

The overall prediction accuracy of our classifier is 90.70% (F1 score) (Table 4). While F1 values are fairly similar across evaluation metrics and films, recall is lower than precision and has a larger standard deviation (7.52%).

In general, high performance correlates with higher numbers of training instances, but not vice versa (Table 4). This concern is moderate as the Pearson correlation coefficient for F1 and the number of instances per fold is -0.65.

Also, precision (94.02%) is higher than recall (87.90%) on average (Table 4), which indicates that $\frac{TP}{TP+FP} > \frac{TP}{TP+FN} \stackrel{\text{def}}{=} FP < FN$ (Equation 7). Since we have the same number of training instances for each class (i.e., $TP + FN = TN + FP$), we can infer that $FP < FN \stackrel{\text{def}}{=} TN > TP$. The results suggest that overall, expert reviews are more accurately predictable than layman reviews. With a further comparison of the standard deviation between precision (4.59%) and recall (7.52%), this finding suggests that expert reviews show lower in-group variability than layman comments. This might be due to professional norms and standards.

The isolated contribution of each feature to prediction accuracy is shown in Table 5 (sorted by decreasing contribution), and the actual values per feature per class are provided in Table 6.

The syntax features have the highest isolated impact, which indicates that out of the considered features, grammar use contributes the most for distinguishing the considered two groups. Looking into POS diversity and parser label diversity as shown in Table 6, expert reviews (0.61 for POS diversity; 0.57 for syntax labels diversity) feature more complex syntax than layman reviews (0.44 for POS

Fold No.	Documentary	Precision	Recall	F1	# Instances
1	SPSZM + TALD	88.36%	75.88%	81.65%	340
2	INJO + PDBTH	85.14%	86.30%	85.71%	146
3	FOIN + PAPR	92.31%	96.00%	94.12%	150
4	GTKER + PL3P	98.18%	94.74%	96.43%	114
5	TCOV + IVWAR	96.55%	91.80%	94.12%	122
6	CTFR + DWAR	96.23%	83.61%	89.47%	122
7	AOKI + HABA	95.65%	75.86%	84.62%	116
8	BLKFSH + HTSAP	100.00%	92.59%	96.15%	108
9	EOTL + HILI	91.80%	94.92%	93.33%	118
10	FBCR + TTDS	96.00%	87.27%	91.43%	110
Average/ Sum		**94.02%**	**87.90%**	**90.70%**	**1446**
Std Dev		4.59%	7.52%	5.15%	/

Table 4: Accuracy from 10-folds cross validation using all features

Feature Type	Feature	Precision	Recall	F1
Syntax	Parts of speech	**92.76%**	**90.01%**	**91.29%**
	Parse tree constituents	85.50%	85.52%	84.45%
Lexical	Transition words	84.74%	69.63%	76.29%
	Entropy	86.32%	68.83%	76.05%
	Unigrams	67.37%	82.75%	73.91%
	Sentiment	73.69%	41.12%	52.30%
Length	Review length	69.63%	74.82%	71.79%
	Avg sentence length	79.29%	64.85%	71.12%

Table 5: Isolated contribution per feature (highest value per column in bold)

diversity; 0.37 for syntax labels diversity). However, this feature may correlate with review length, which is considerably higher for experts (362 words) than for laymen (107 words). Also, experts use more nouns than laymen, while laymen use more verbs and adjectives.

The choice of salient words (unigrams) has a strong impact on recall, while text informativeness (entropy) and transition words rather contribute to precision. Given the aforementioned definition of precision and recall, this result hints at some uniformity or consistency of word choice in laymen reviews, and at higher vocabulary diversity as well as more coherent structure in expert reviews. Sentiment is the weakest contributor to the prediction.

Further analyzing the differences between both groups (Table 6), we find that experts, in comparison to laymen, write longer reviews and longer sentences, use more complex syntax, provide more new information (i.e., expert reviews have higher entropy values than layman reviews), have a higher diversity in their vocabulary, and use fewer emotional words. Some of these features might correlate with review length, but overall, these findings might be explainable by a professional text production style that reflects established norms and rules of journalistic writing. Based on our data, short reviews are the strongest defining feature for non-expert reviews. Layman reviews are also more opinionated and emphasize points made more strongly (i.e., high and fluctuating Sentiment%).

In addition to these quantitative analyses, we conducted a qualitative analysis by reading through the top 20 unigrams (based on TF*IDF) for each film to better understand difference in content and writing between experts and laymen. Table 7 provides an illustrative example for two randomly selected films, and we refer to this example in the following discussion of descriptive features per category. Overall, we find that experts frequently refer to 1) people involved in making and producing films ("director"; "morgan," "spurlock"), 2) film titles ("inside," "job"), 3) cinematographic concepts ("moore" as *Michael Moore style*), and 4) awards and festivals. Also, experts connect issues addressed in films to current affairs and higher level topics ("obesity"), and provide details or background information (e.g., "hubbard"; who frequently appeared in expert reviews of INJO, represents *Glenn Hubbard;* an economist who previously worked for the federal government). Expert reviews entail specific concepts ("obesity") and formalities ("Mr."), while laymen use more casual terms ("fat"; "bad"). Laymen reviews represent substantial engagement with the topic of a film ("eat"; "diet"), contextualize issues in peoples' regular lives ("people"; "day"; "money"; "job"; "school"; "healthy"), and contain more subjective terms ("good"; "bad").

Feature	Expert (AVG±STD)	Layman (AVG±STD)
Entropy	2.94±0.79	1.86±0.69
Sentiment%	0.09±0.03	0.12±0.12
Transition%	0.05±0.02	0.07±0.05
Ratio addition	0.63±0.24	0.44±0.33
Ratio example	0.02±0.07	0.01±0.05
Ratio emphasis	0.04±0.07	0.12±0.21
Ratio concession	0.15±0.17	0.10±0.18
POS diversity	0.61±0.16	0.44±0.19
Ratio NN	0.36±0.06	0.25±0.12
Ratio VB	0.14±0.04	0.18±0.07
Ratio JJ	0.09±0.03	0.11±0.12
Ratio RB	0.04±0.02	0.06±0.06
Syntax label diversity	0.57±0.17	0.37±0.19
Review length	362.00±421.56	107.50±188.65
Avg sentence length	27.52±13.88	15.16±8.66

Table 6: Values and variance of features per class

5.2 Error Analysis

The confusion matrix (Table 8) shows that our classifier predicts expert reviews with higher accuracy (93.36%) than laymen reviews (86.17%). More importantly, laymen are more likely to be mistaken for experts (13.83%) than

SPSZM	Expert	size, spurlock, mcdonald, super, film, year, days, director, big, moore, month, obesity, company, million, morgan, day, people, festival, burger, american
	Layman	mcdonalds, people, spurlock, movie, film, diet, mcdonald, eat, day, fat, make, school, healthy, bad, time, eating, good, watch, body, experiment
INJO	Expert	ferguson, inside, film, job, charles, crisis, director, men, mr, company, financial, global, banks, documentary, bankers, economic, end, hubbard, street, crash
	Layman	film, movie, wall, people, government, street, job, money, documentary, banks, great, financial, crisis, inside, world, watch, good, american, loans, made

Table 7: Top 20 Unigrams for Case Study

vice versa (6.64%). Looking into laymen reviews labeled as expert reviews, we find that these texts are long, detailed, and contain subject matter expertise. Expert reviews that got misclassified as laymen reports were typically short. These findings further substantiate our previously made point that some laymen write expert-level reviews, and vice versa.

To analyze our prediction errors more in depth, we selected a random sample (N=62) of misclassified reviews from both classes. We removed the class labels from these documents and asked two independent human annotators to code the texts as expert or layman reviews. Their inter-coder agreement was 45.2%, which suggests that categorizing these cases is also hard for humans.

We further discussed the label assignments with our two human coders. Trends emerging from their observations and our discussion are summarized in Table 9, where we synthesize the humans' feedback into a high or low value per identified feature and class. The features and values that the humans identified strongly overlap with those considered for supervised learning, e.g., level of detail (high for experts), subject matter expertise (high for experts), emotionality (high for laymen), and formal (experts) versus informal (laymen) writing styles. Beyond that, the close reading analysis also revealed additional features, e.g., differences in the usage of personal pronouns ("I" for laymen) and comparatives and superlatives (high for laymen), which can be used in future work, e.g. as new features. Overall, the majority of cases where both the classifier and the humans were incorrect are short expert reviews.

		Prediction	
		Expert	Layman
Truth	Expert	93.36%	6.64%
	Layman	13.83%	86.17%

Table 8: Error analysis

6 Discussion, Conclusions and Limitations

We have developed a binary classifier that predicts whether a review was authored by an expert or a layman with an accuracy of (90.70% (F-measure). Our work is novel with respect to its goal, focus, and potential applications. While prior work has focused on predicting commercially motivated

Features identified by human coders	**Expert**	**Layman**
Deep analysis including identification of different opinions about a given topic	High	Low
Technical details, e.g. running time, and screenings references, such as film festivals and award nominations	High	Low
Movie jargon, subject matter expertise about film-making ("guerrilla filmmaking style")	High	Low
Words and short phrases with strong emotions ("I strongly recommend this film to any ocean lover")	Low	High
Comparatives and superlatives ("Charles Ferguson made the best documentary")	Low	High
Personal pronouns used as self-reference to reviewer ("I")	Low	High
Questions to convey disbelief ("What about the effort to find the person, or people who did do it?")	Low	High
Casual, informal style ("In the second doc").	Low	High
Words with all letters in upper case ("Watch this film NOW").	Low	High
Smaller variability in vocabulary (e.g., duplicated words/phrases)	Low	High
Grammatical errors, sloppiness	Low	High

Table 9: Manually identified features and values

features, e.g. the rating, helpfulness, and sentiment of reviews, we aim to predict the type of author. We believe that this work enhances our understanding of the impact of issue-focused media, in our case documentary films, on different types of users.

Our results suggest that experts write comparatively lengthier and more detailed reviews with more complex grammar, higher entropy, and lower emotionality. Laymen are less object (noun) and more action (verb) oriented, and engage more emotionally with the content of a film. The relevance of these features was empirically demonstrated, and then manually verified and extended by human judges.

The generalizability of our findings is limited by several choices we made: First, we worked with data from two particular sources, i.e., expert reviews published mainly in major newspapers and retrieved from LexisNexis Academic, and laymen reviews collected from Amazon. Although our data come from different platforms, we argue that the considered text features are not a function of the type of source, but of the way in which experts versus laymen express their impressions of a film. Second, the first choice furthermore entails the assumption that user-generated reviews on Amazon are authored by laymen, while professional writers author expert reviews. We have shown that this assumption does not always hold: For the case of erroneous predictions, laymen are about twice as likely to be identified as experts than vice versa. Third, we also tried to use additional sources of reviews, but were constrained by the terms of service for these sites (e.g., Metacritic, Rotten Tomatoes). Fourth, since our primary goal is feature selection and analysis, we report results based on only one learning algorithm, namely SVM. In the future, we plan to explore the contribution of our binary classifier as a feature for predicting review ratings, helpfulness and authenticity. We will also test if prediction accuracy can be further increased by adding features that were detected by our human annotators during the error analysis process, namely the consideration of personal pronouns, comparatives and superlatives.

Finally, even though it is peripheral to the NLP work presented in this paper, looking at our results from a social or media impact assessment perspective, we find that laymen do engage with the content of a film, and contextualize issues raised in documentaries in their personal lives. These effects indicate public awareness and impact on information consumers. Our work might help researchers and practitioners in the field of impact assessment to understand how different groups of stakeholders reflect on a topic or a work of art (Barrett & Leddy, 2008; Chattoo & Das, 2014; Clark & Abrash, 2011; Diesner et al., 2014; Green & Patel, 2013; John & James, 2011; Napoli, 2014).

Acknowledgement

This work was supported by the FORD Foundation, grant 0155-0370, and by a faculty fellowship from the National Center of Supercomputing Applications (NCSA) at UIUC. We are also grateful to Amazon for giving us permission to collect reviews from their website. We also thank Sandra Franco and Harathi Korrapati from UIUC for their help with this paper.

Reference

Amblee, N., & Bui, T. (2007). Freeware downloads: An empirical investigation into the impact of expert and user reviews on demand for digital goods. In *Proceedings of the Americas Conference on Information Systems, AMCIS*, Paper 21.

Barrett, D., & Leddy, S. (2008). Assessing creative media's social impact. The Fledgling Fund.

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing, 67*(4), 103-117.

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

Chattoo, C. B., & Das, A. (2014). Assessing the social impact of issues-focused documentaries: Research methods & future considerations. Center for Media and Social Impact, School of Communication at American University.

Clark, J., & Abrash, B. (2011). Social justice documentary: Designing for impact. Center for Social Media, School of Communication at American University.

de Albornoz, J. C., Plaza, L., Gervás, P., & Díaz, A. (2011). A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating. *Advances in Information Retrieval* (pp. 55-66). Berlin Heidelberg: Springer.

de Jong, I. K., & Burgers, C. (2013). Do consumer critics write differently from professional critics? A genre analysis of online film reviews. *Discourse, Context & Media, 2*(2), 75-83.

De Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of International Conference on Language Resources and Evaluation, LREC*, (pp. 449-454), (Vol. 6).

De Marneffe, M.-C., & Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Stanford University.

Diesner, J., Kim, J., & Pak, S. (2014). Computational impact assessment of social justice documentaries. *Journal of Electronic Publishing, 17*(3).

Diesner, J., Rezapour, R., & Jiang, M. (2016). Assessing public awareness of social justice documentary films based on news coverage versus social media. In *Proceedings of the iConference*.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2011). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-27.

Flanagin, A. J., & Metzger, M. J. (2013). Trusting expert-versus user-generated ratings online: The role of information volume, valence, and consumer characteristics. *Computers in Human Behavior, 29*(4), 1626-1634.

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498-1512.

Green, D., & Patel, M. (2013). Deepening engagement for lasting impact; A framework for measuring media performance and results. John S. and James L. Knight Foundation and Bill & Melinda Gates Foundation.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining, SIGKDD*, (pp. 168-177), (Vol. 04), ACM.

Jiang, M., & Diesner, J. (2016). Issue-focused documentaries versus other films: Rating and type prediction based on user-authored reviews. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, (pp. 225-230), ACM.

Jindal, N., & Liu, B. (2007). Review spam detection. In *Proceedings of the 16th International Conference on World Wide Web, WWW*, (pp. 1189-1190), ACM.

Jindal, N., Liu, B., & Lim, E.-P. (2010). Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM*, (pp. 1549-1552), ACM.

John, S., & James, L. (2011). Impact: A practical guide for evaluating community information projects. Knight Foundation.

Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP*, (pp. 423-430), Association for Computational Linguistics.

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., & Yu, H. (2010). Structure-aware Review Mining and Summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING*, (pp. 653-661).

Liu, Y., Huang, X., An, A., & Yu, X. (2008). Modeling and predicting the helpfulness of online reviews. In *Proceedings of the IEEE International Conference on Data Mining, ICDM*, (pp. 443-452).

Mackiewicz, J. (2009). Assertions of expertise in online product reviews. *Journal of Business and Technical Communication, 24*(1), 3-28.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics, 19*(2), 313-330.

McAuley, J., & Leskovec, J. (2013). From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *In Proceedings of the 22nd international Conference on World Wide Web, WWW*, (pp. 897-908), ACM.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon. com. *MIS quarterly, 34*(1), 185-200.

Napoli, P. (2014). Measuring media impact: An overview of the field. Media Impact Project, USC Annenberg Norman Lear Center.

Nelson, P. (1970). Information and consumer behavior. *The Journal of Political Economy, 78*(2), 311-329.

Nelson, P. (1974). Advertising as information. *The Journal of Political Economy, 82*(4), 729-754.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *43rd Annual Meeting on Association for Computational Linguistics, ACL*, (pp. 115-124), Association for Computational Linguistics.

Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do experts and novices evaluate movies the same way? *Psychology & Marketing, 26*(5), 470-478.

Rezapour, R., & Diesner, J. (2017). Classification and Detection of Micro-Level Impact of Issue-Focused Films based on Reviews. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, ACM.

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*, (pp. 417-424), Association for Computational Linguistics.

Weaver, W., & Shannon, C. E. (1949). The Mathematical Theory of Communication. *University of Illinois Press*. Urbana, Illinois.

Wilson, T., Wiebe, J., & 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, EMNLP*, (pp. 347-354), Association for Computational Linguistics.

Wu, G., Greene, D., & Cunningham, P. (2010). Merging multiple criteria to identify suspicious reviews. In *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys*, (pp. 241-244), ACM.

Yang, Y., Yan, Y., Qiu, M., & Bao, F. S. (2015). Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL and the 7th International Joint Conference on Natural Language Processing, IJCNLP*, (pp. 38-44), (Vol. 2: Short Papers).

Zhang, R., Yu, W., Sha, C., He, X., & Zhou, A. (2015). Product-oriented review summarization and scoring. *Frontiers of Computer Science, 9*(2), 210-223.

Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM*, (pp. 43-50), ACM.