Identifying the Approach to Movie Reviews using Natural Language Processing

David Simmonds[a)] and Abhinandan Chowdhury[b)]

Savannah State University, Savannah, Georgia 31404, USA.

[a)]*simmondsd@savannahstate.edu.*
[b)]*Corresponding author: chowdhury@savannahstate.edu.*

Abstract. Everyday viewers now use social media platforms to express opinions on movies. Therefore if someone wants to watch a movie, he can now leverage thousands of reviews including ratings and comments on websites such as *Rotten Tomatoes* and IMDb. Comments are valuable if reviewer and reader have common focus in their viewing habits. In this paper, we examine whether users can intuit the focus of movie reviews in 6 dimensions: Human interest, Recommendations given, Extremity, Technicality, Comparativeness and Plot Summary. We find strong evidence to suggest that viewers can indeed connect with reviewers in some of these dimensions. We contribute to the literature by offering another way to approach movie review categorization by using *Natural Language Processing* (NLP).

INTRODUCTION

We are no longer limited to the selection of movies available from DVD stores like Blockbuster. Streaming sites such as Netflix, Amazon Prime and Hulu now provide a massive collection of movies which can be streamed on demand, resulting in a plethora of options which may suit one's viewing profile. The downside is that viewers now suffer from information overload. Dichter [1] established through multiple interview techniques that movie watching would be affected by a friend or relative who provides opinions and feelings about a movie. This is the basis of the rise of online ratings, termed electronic Word of Mouth (eWom), towards becoming the new way of finding movies. Like traditional reviews, online reviews consist of two parts. The first is a numerical rating for the movie, for example, a number in the scale of 1 to 10 each review/reviewer gives on IMDb. The second is a narrative aimed at expressing the details of the movie which makes it appealing or not, to the reviewer. The former is meant as an objective score to be averaged, while the latter serves to justify the viewer's review in a subjective manner.

While online reviews are not without flaws, consumers do put a lot of information into them, describing their experience, making it relatable to the reader, and possibly more helpful. Typically the act of reviewing is motivated by self-confirmation, and stems from the desire to help the community, through self-expression [1]. As a result, since streaming sites have now become the staple of at-home viewing, the eWom of ordinary viewers have started being valued more than the reviews coming from the professional movie reviewers and critics which sometimes can come across as "highfalutin" and out of touch. That is why, online reviews have taken prominence in consumer decisions because they are seen as more realistic as if they are coming from their peers.

One advantage of movie reviews is that they do not constrain the reviewer to a restrictive format which would limit their expressiveness. Viewers have access to thousands and sometimes hundreds of thousands of these personalized narratives advising them whether or not to watch a movie. Processing this deluge of reviews would be a painstaking process, and certainly not worth the effort. So while it is true that narratives in movie reviews add dimensionality to the numeric rating, at the same time there exists disadvantages as well while using online reviews. One of them being the lack of template which would quickly guide the reader to the parts of the review they are most interested in. Additionally since these are the opinions expressed by people in the broader society, there is the deeper subjectivity and less uniform approach taken by lay-people in any area [2]. With everyone vying for relevance and a voice on the web, readers who are short on time and more interested in experiencing the movie than analyzing it, the eruption of reviews can easily become too much information [3].

To help with this information overload, text mining goes a long way to extracting information that is not only relevant to the current movie, but also in recommending other movies [2]. It is possible for humans to determine where a review lies, along several spectrum such as its emotional content, or extremity of expression – as indicated by our data set and results. But making the movie selection process effective is still laborious, and hence inadequate for humans [4].

Hence the rise of natural processing techniques which attempt to help users to make sense of movie review process. Computerized techniques like Natural Language Processing (NLP) not only help viewers to manage the deluge of online reviews, but also help to overcome the risk factor in the mind of every purchaser [5]. In this paper, we focus

on features of the narrative component. Especially we examine whether readers notice and identify the tone of a review in 6 dimensions we have proposed, in addition to the information such as the number of rating stars, or general helpfulness. Our aim is to propose an approach to help users seek out such movie reviews which would subsequently enable them to experience more personalized entertainment by watching movies which are in sync with their emotional maps – dimensions in which they appreciate movies [6].

We use classification with 2 polarities in each dimension. This approach which is recommended in [2] for its accuracy, is based on pre-training. In the real world, there are thousands of reviews to make sense of, and there is no way it is possible for the average user to know which are the most valuable [3], without using NLP. In this paper, we make use of a human-tagged data set of movie reviews to derive more ways to approach review categorization. We analyze 1800 movie reviews tagged by 3 persons who represent regular movie goers. Their efforts represent ordinary user's attempts to extract an overall impression of a movie from the tens of thousands of reviews they encounter [2].

To prove our hypotheses using NLP, we examine the agreement of our tagged data set, with the classification of Naïve Bayes, Multinomial Naïve Bayes (MNB) and Bernoulli Naïve Bayes (BernoulliNB) based on a lexical bag-of-words [7]. Using this collection of techniques, we aim to show that users easily detect the slant on movie reviews along the dimensions of *Technicality*, *Recommendations given*, *Extremity*, *Human interest*, *Comparativeness* and *Plot Summary*.

THEORY

Societal Impact of Social Media

Zúñiga et al. [8] have shown that social media increase social participation in topics of interest in society. The societal reach that social media can have is beyond question, as Barack Obama demonstrated in winning the presidency with the help of Social Media [9], and the organization of protests in Egypt using Twitter [10]. Facebook users have also used Social Networking to successfully organize pro-immigration marches [11]. On the topic of movies, Duan et al. [12] found that the number of online reviews about a movie generates the awareness among the intended audiences – a signal that the movie is popular and has mass market appeal, even if it is not critically acclaimed.

Social Media Reviews: The new Word of Mouth

Undoubtedly there is value in online reviews as a method of transmitting knowledge about product value, across the breadth of consumers [12], with millions seeking guidance by using the thousands of movie reviews. Online reviews are the new form of eWom as a means of spreading information on products and online reviews via social media [13]. This transmission of information has demonstrated its power, especially when products or companies attract negative sentiment [14]. Terms used online to describe the negative reactions to such movies have included words or phrases such as "storm," "internet backlash," "broke the internet," and "internet outrage."

The needs of the one

While reviews have gone online and though the new format was supposed to provide transparency through access to aggregated, instant feedback, it does not mean that individual reviewers would refrain from attempting to discredit a movie through a single voice [9], which has prompted streaming platforms to attempt strong moderation of comments to avoid inflammatory statements.

The "helpfulness" metric, of which Amazon was one of the first platforms to introduce, is a step toward filtering reviews to promote the least biased reviews. They assist readers in quickly determining the most trustworthy reviews, free from extreme reactions from any one reviewer. These helpful reviews tend to be thoughtfully generated by describing, in a balanced way, strengths and weaknesses of the movie. Additionally, like many other social media platforms, they allow readers to look at the reviewers' expertise as reflected in their review history. This also allows reviewers to establish their standing within the community of consumers, as someone who provides well-thought-out reviews as opposed to knee-jerk reactions.

The needs of the many – eWOM

The way reviews are perceived by readers, and subsequently affect readers' intention to action, is affected by the way readers typically use information in making purchases. Different receivers of social media reviews react very differently depending on their prior knowledge, their level of engagement, perspectives and other products they have had experience with [13]. The trust they have in reviews they have come across also moderates the influence of the reviews on their intentions to watch a movie. For example, the anonymity of reviewers using a handle instead of their real name, affects their perceived trustworthiness and knowledgeability/expertise [13].

The value of movie watching time

Despite their ability to amalgamate online seller reputations, there is still a risk factor in the mind of the viewer, that eWOM does not totally overcome. It was argued in [5] that when taking word of mouth into consideration, typically decisions are made based on two factors. The first is the risk factor of spending resources such as time without getting value – in this case entertainment from a movie. The other factor is the prospect of watching a better movie, especially one that that is in sync with the viewer's preferred type of movies. The likelihood of risk aversion will be greater if the buyer prefers to watch in a movie theater [5]. Therefore, if a viewer already has his heart set on watching a particular movie, then negative reviews are more likely to be ignored or discounted unless they incorporate compelling evidence or arguments [15].

Status within the consumer community

Consumers have demonstrated the tendency to actively spread word-of-mouth of a positive nature, motivated by the desire to help fellow consumers and also in order to make themselves look and feel good about their expertise on display. On the other hand, spreading negative sentiment by word-of-mouth seems to be rooted in a need to exact a sense of justice for their negative experience [16].

A bigger level of agreement between the reviewer and the rest of the community is a psychological stimulus to the reviewer since it means that the reviewer represents the popular sentiment which further legitimizes their feelings about a movie. This motivates the reviewer to spread that word-of-mouth even more to reinforce their sense of expertise [16].

METHOD

According to Sun et al. [3], Naïve Bayes classifier on which the other classifiers are based is a widely implemented classification algorithm. They are fed a corpus of data known as training set with two polarities – after which they are able to classify additional documents. In this case our testing set is divided into one of the two categories. We use classifiers to categorize reviews in the 6 dimensions: Emotion vs. Technical, Extremity of expression, Recommendations focused, Comparison to other movies, and technical vs. human aspects of the movie.

Hypotheses

H1: Readers are able to instinctively determine whether the focus of a review is emotional or technical.

Detecting emotion in social media is useful when it comes to finding the preferences of potential viewers [17]. They used principal-based approach (PBA) which seeks to determine the emotional connectedness between readers and viewers in online reviews. They were able demonstrated that there was deep connection using the system. Chang et al. [18] created emotion maps of readers using an software they developed. It allows movie recommender systems to compare the emotional maps between movies and viewers.

H2: Readers can determine whether the reviewer is focused on the human interest value of a movie in contrast to its raw entertainment features.

Parks [19] demonstrated a growing divergence between human interest and factual journalism. They showed that journalist's ability to distinguish facts from human interest allowed them to infuse relevant human interest in order to make their reporting engaging. Gans [20] recognized human interest as one of 6 types of story telling which can be interesting. Parks [19] showed that the human interest aspect of a story affects readers' perspectives and whether they could remember it.

H3: Readers are able to differentiate between reviews that stand on their own, in contrast to reviews that are comparative.

Movie reviews about sequels tended to be more likely to affect a viewer's derived pleasure when the number of the sequel is mentioned compared to a named sequel. Readers spent more time on reviews using the number of the sequel in the title. They also remember more about the sequel and also had more favorable attitudes toward them [21].

H4: Readers can tell the difference between a review based on plot summaries versus reviews that are generally less structured.

A plot summary conveys much more information about a movie. Hoang [22] used multiple machine learning models to demonstrate that plot summaries do impact decisions concerning movies. They created *Word2vec* which quickly separate the positive and negative comments about a movie. They were able to create summaries which were quicker for readers to scan and then decide whether to watch a movie. Khan et. al. [23] developed a movie review summarizer which classified words in a review based on their major features.

H5: Readers can tell the difference between a review which offers recommendation as against ones that do not.

When considering reviews, readers tend to be more convinced by reviews which use stronger endorsement. An explicit recommendation to purchase a product or service is more impactful than a non-committal expression of how much it was liked or disliked. As Packard et al. have shown in [24], a reviewer who says "you have to watch this movie" will be taken more seriously than a reviewer who says "I really loved this movie." Experienced reviewers tend to give more explicit recommendations indicating that the potential viewer should or should not view a movie. Additionally, it was explored in [25] whether readers would deem explicit recommendations as being more valuable. Their results gave strong support to their hypothesis.

H6: Readers are attuned to the level of dispassion or extreme expression in a movie review.

It was discovered in [26] that emotion in a negative review not only de-emphasizes its informational feature, it also mitigates the negative effect on the reader because readers see them as less useful. It was concluded in [6] that the review of a particular movie conveys an overall higher level of good or bad feelings from reviewers. This was palpable to the reader of these movie reviews.

DATA ANALYSIS

A total of 1800 movie reviews from 12 movies were downloaded from Amazon Prime to create a spreadsheet template. Each row of the spreadsheet contains the movie title, numeric rating which was hidden and the review text. The spreadsheet also included a column representing each of the 6 dimensions. Each column was modified to include a drop-down list from which the value of each dimension was picked. 3 persons were employed to categorize each movie along all 6 dimensions. They tagged the movies along the following dimensions: emotional vs. technical, main theme focused on human interest or entertainment, comparison to other movies or the movie is reviewed on its own merit, plot summary provided or not, recommendations focused or not, and extremity of expressions. It is understandable that each dimension is divided into two categories. For example, in the case of the last dimension, namely *extremity of expressions*, one category would be the type of reviews in which the reviewers strongly convey their feelings by using highly persuasive words or phrases, whereas other category could include the types of reviews of which the tone is somewhat dispassionate.

In the pre-processing stage of the program written in Python 3.0, each word in the document set is stemmed using Natural Language Toolkit's (NLTK) Porter Stemmer [27]. It means that each word is replaced by its root word. For example, "banker," "banks" and "banking" are replaced by "bank." We have removed *Stop words* which add minimal information to a sentence but make it sound smoother e.g., "a," "it," "in," "the" etc. Punctuation is also removed. Each document is converted to a bag of words. A list is then created which contains each occurrence of every word on each

document. From this, a frequency distribution is created with the most frequently occurring words (and the number of occurrences across all documents) at the top. The top n words (1000 in our case) are then used to create a *feature universe*. A document featureset is then created for each document, by comparing it to the feature universe. Each document featureset is a list of n binary values. A TRUE in the n-th position of the document featureset means that the n-th word in the feature universe exists in the original document.

A python program is written to facilitate the following tasks:

- Each review was word tokenized to create a bag of words.
- From this bag of words, three featuresets were alternately extracted. These featuresets were raw-words, no-stop words and word-stems.
- From each featureset, a sparse matrix was created to be fed into the classifier.
- Each of the 3 featuresets was classified and tested using Naïve Bayes, Multinomial Bernoulli (MNB), and Bernoulli Naïve Bayes (BernoulliNB) [28].

They work in two phases, first of which is the training phase. A training data set consists of hundreds of document featuresets which are pre-tagged by human beings into two categories for each dimension. Each document featureset is a tuple, the first of which is a collection of features known as a "bag of words." Remember though, that each "word" is really a binary value indicating if it is in the feature universe. The second part of the tuple is a number denoting one of the two predetermined categories. The training set is fed to the classifier, which learns by making a list of the most informative words in each category. For example, a word is considered one of the most informative words for category A, if it has a high ratio of occurrence in category A versus its occurrence in category B. Hence, the most informative words are the words occurring most frequently in either category, compared to another.

The test data set also has hundreds of document featuresets with predetermined categories. The second phase, known as the testing phase (which uses the test set) is further broken down into two activities. The first activity includes for each document featureset in the test set, the classifier checks the Bayesian probability that the featureset is in either category. It is performed by looking at the most informative words and based on the sum of the weighted ratios of existence in the category, it selects the category with the greater probability. In other words, if the document has a combination of the most popular words associated with a category, it will make an informed "guess" that the document belongs to that category.

In the second activity, the classifier compares each test document's pre-tagged category with the category it had assigned it to. It then computes its own accuracy as the proportion of accurately "guessed" categories over the total number of documents in the test set.

RESULTS

Each table consists of 4 columns as described below:

- Word: One of the words used to categorize documents. It appeared in the top 1000 words used in the Feature Universe to categorize documents. These 25 words are the most useful words (most informative features) out of the 1000 selected overall.
- Y : N: Whether the word was found in the category indicated by the table-heading (for e.g., "Technicality"). $1:0$ indicates that it was found more often in the (technical) category. $0:1$ indicates it was found more often in the other (non-technical) category.
- Ratio: Ratio of times the word was found in the category compared to the number of times the word was found in the opposite category. For example, Y : N of $1:0$ labelled with a ratio of $7.7:1.0$ means it was found 7.7 times more frequently in the category than it was found in the opposite category.
- Significance: What the word intuitively represents in the movie database. They were tagged by the researchers.

H1 posited that readers are able to instinctively determine whether the focus of a review is emotional or technical. Below is the accuracy of the top 3 classifiers in categorizing the reviews according to the Technicality feature. From the results we can see that **H1** has strong support. Whether the actors were discussed appeared to have a slight impact on whether the review was non-technical (category 0).

TABLE I. Technicality: Most Informative Features

Word	Y : N	Ratio	Significance
program = True	1 : 0	7.7 : 1.0	
garbag = True	0 : 1	7.2 : 1.0	
stupid = True	0 : 1	6.8 : 1.0	
alreadi = True	0 : 1	6.3 : 1.0	
adult = True	1 : 0	6.2 : 1.0	
uplift = True	1 : 0	6.2 : 1.0	
favorit = True	0 : 1	6.0 : 1.0	
mckinnon = True	0 : 1	5.4 : 1.0	Actor
simpli = True	0 : 1	5.4 : 1.0	
sort = True	0 : 1	5.4 : 1.0	
wow = True	0 : 1	5.4 : 1.0	
delight = True	1 : 0	5.2 : 1.0	
girlfriend = True	1 : 0	5.2 : 1.0	
either = True	0 : 1	4.9 : 1.0	
etc = True	0 : 1	4.9 : 1.0	
hate = True	0 : 1	4.9 : 1.0	
hemsworth = True	0 : 1	4.9 : 1.0	Actor
kate = True	0 : 1	4.9 : 1.0	Actor
cameo = True	0 : 1	4.9 : 1.0	
panther = True	0 : 1	4.9 : 1.0	Movie Title
worst = True	0 : 1	4.7 : 1.0	
sound = True	1 : 0	4.7 : 1.0	
oh = True	0 : 1	4.4 : 1.0	
suck = True	0 : 1	4.4 : 1.0	
suppose = True	0 : 1	4.4 : 1.0	

TABLE II. Accuracy of the top 3 classifiers in categorizing the reviews according to the Technicality feature.

Reviewer Focus	Featureset	NB	MNB	BernoulliNB
Technicality	raw-words	90	80	90
Technicality	word-stem	88	80	86
Technicality	words-nostop	92	88	94

H2 posited that readers can instinctively determine the central focus of reviews. Below is the accuracy of the top 3 classifiers in categorizing the reviews according to the major focus. From the results we can see that **H2** has weak support. When the actors are discussed, it can be easily argued that the review was focused on the human-interest aspect of the movie (category 0). What is interesting is that the reviewers tend to discuss the characters using the actor's name in lieu of the actual character. This is not surprising since we tend to think of the role played by the actor instead of the actual character.

TABLE III. Main Theme: Most Informative Features

Word	Y : N	Ratio	Significance
washington = True	0 : 1	12.7 : 1.0	Actor
audrey = True	0 : 1	9.6 : 1.0	Actor
altern = True	1 : 0	7.8 : 1.0	
steal = True	0 : 1	7.7 : 1.0	
annoy = True	0 : 1	7.1 : 1.0	
beyond = True	1 : 0	6.8 : 1.0	
normal = True	0 : 1	6.5 : 1.0	
regina = True	0 : 1	6.5 : 1.0	Actor
tautou = True	0 : 1	6.5 : 1.0	Actor
dialog = True	1 : 0	6.1 : 1.0	Actor
video = True	1 : 0	6.1 : 1.0	Actor
tiffani = True	0 : 1	6.0 : 1.0	Actor
hall = True	0 : 1	5.9 : 1.0	
latifah = True	0 : 1	5.9 : 1.0	Actor
trailer = True	0 : 1	5.9 : 1.0	
trip = True	0 : 1	5.9 : 1.0	
costum = True	1 : 0	5.8 : 1.0	
queen = True	0 : 1	5.8 : 1.0	Actor
haddish = True	0 : 1	5.7 : 1.0	Actor
denzel = True	0 : 1	5.7 : 1.0	Actor
pratt = True	0 : 1	5.4 : 1.0	Actor
foreign = True	1 : 0	5.4 : 1.0	
later = True	1 : 0	5.4 : 1.0	
perfectli = True	0 : 1	5.3 : 1.0	
wors = True	1 : 0	4.8 : 1.0	

TABLE IV. Accuracy of the top 3 classifiers in categorizing the reviews according to the major focus.

Reviewer Focus	Featureset	NB	MNB	BernoulliNB
Main Theme	raw-words	63.1	62.2	63.9
Main Theme	word-stem	69.3	68.9	70.1
Main Theme	words-nostop	63.1	62.2	65.6

H3 posited that readers are able to differentiate between reviews that stand on their own, in contrast to reviews that are comparative. Below is the accuracy of the top 3 classifiers in categorizing the reviews according to the Movie-Comparison feature. From the results we can see that **H3** is not supported. Reference to the other movies in the movie franchise (e.g., prequels, sequels and the cinematic universe) are naturally part of movie comparisons.

TABLE V. Compared: Most Informative Features

Word	Y : N	Ratio	Significance
pitch = True	1 : 0	12.9 : 1.0	
place = True	1 : 0	11.0 : 1.0	
previou = True	1 : 0	11.0 : 1.0	Refer: Franchise
sequel = True	1 : 0	10.6 : 1.0	Refer: Franchise
die = True	1 : 0	10.0 : 1.0	
add = True	1 : 0	9.0 : 1.0	
annoy = True	1 : 0	9.0 : 1.0	
destroy = True	1 : 0	8.1 : 1.0	
fail = True	1 : 0	8.1 : 1.0	
jedi = True	1 : 0	8.1 : 1.0	
remak = True	1 : 0	8.1 : 1.0	Refer: Franchise
ghost = True	1 : 0	8.0 : 1.0	
compar = True	1 : 0	7.7 : 1.0	Refer: Franchise
origin = True	1 : 0	7.7 : 1.0	Refer: Franchise
ghostbust = True	1 : 0	7.4 : 1.0	
alon = True	1 : 0	7.1 : 1.0	
fresh = True	1 : 0	7.1 : 1.0	Refer: Franchise
hold = True	1 : 0	7.1 : 1.0	
luca = True	1 : 0	7.1 : 1.0	
near = True	1 : 0	7.1 : 1.0	
refer = True	1 : 0	7.1 : 1.0	Refer: Franchise
univers = True	1 : 0	7.1 : 1.0	Refer: Franchise
war = True	1 : 0	7.0 : 1.0	
american = True	0 : 1	6.9 : 1.0	
cameo = True	1 : 0	6.7 : 1.0	Refer: Franchise

TABLE VI. Accuracy of the top 3 classifiers in categorizing the reviews according to the movie comparison feature.

Reviewer Focus	Featureset	NB	MNB	BernoulliNB
Comparison	raw-words	50	57.3	47.8
Comparison	word-stem	52.8	57.9	50.0
Comparison	words-nostop	51.1	54.5	47.2

H4 posited that readers can tell the difference between a review based on plot summaries versus reviews that are generally less structured. Below is the accuracy of the top 3 classifiers in categorizing the reviews according to the Plot-Summary feature. From the results we can see that **H4** has strong support. Unsurprisingly, movie names appear frequently in plot summaries. Also genre and what the movie was about naturally appear as well.

TABLE VII. Summarized-Plot: Most Informative Features

Word	Y : N	Ratio	Significance
space = True	1 : 0	17.6 : 1.0	Movie Title
black = True	1 : 0	16.7 : 1.0	Movie Title
three = True	1 : 0	9.1 : 1.0	Movie Title
white = True	1 : 0	8.3 : 1.0	
inspire = True	1 : 0	7.2 : 1.0	Genre
rather = True	1 : 0	7.2 : 1.0	
until = True	1 : 0	7.2 : 1.0	
remak = True	0 : 1	6.7 : 1.0	References
him = True	1 : 0	6.6 : 1.0	
may = True	1 : 0	6.2 : 1.0	
mean = True	1 : 0	6.2 : 1.0	
suspens = True	1 : 0	6.2 : 1.0	Genre
war = True	1 : 0	6.2 : 1.0	Plot
won = True	1 : 0	6.2 : 1.0	
world = True	1 : 0	6.2 : 1.0	
enjoy = True	0 : 1	5.5 : 1.0	
melissa = True	0 : 1	5.4 : 1.0	Actor
danc = True	1 : 0	5.2 : 1.0	
detail = True	1 : 0	5.2 : 1.0	
doe = True	1 : 0	7.5 : 2.0	
everyth = True	1 : 0	5.2 : 1.0	
hype = True	1 : 0	5.2 : 1.0	
order = True	1 : 0	5.2 : 1.0	
overcom = True	1 : 0	5.2 : 1.0	
effect = True	0 : 1	4.9 : 1.0	

TABLE VIII. Accuracy of the top 3 classifiers in categorizing the reviews according to the plot-summary feature.

Reviewer Focus	FeatureSet	NB	MNB	BernoulliNB
Summarized-Plot	raw-words	93.2	93.2	88.1
Summarized-Plot	word-stem	89.8	89.8	86.4
Summarized-Plot	words-nostop	88.1	83.1	79.7

H5 posited that viewers can tell the difference between a review which offers recommendation as against ones that do not. From the results we can see that **H5** has strong support. Recommendation focused terms such as “recommend" and “must see" were used by the classifier to identify the category. Chris Pratt seems to be unrelated to movie recommendations.

TABLE IX. Recommended: Most Informative Features

Word	Y : N	Ratio	Significance
recommend = True	1 : 0	21.4 : 1.0	Recommendation
highli = True	1 : 0	16.3 : 1.0	
must = True	1 : 0	9.7 : 1.0	Recommendation
children = True	1 : 0	8.1 : 1.0	
english = True	1 : 0	8.1 : 1.0	
space = True	1 : 0	7.7 : 1.0	
pratt = True	0 : 1	7.2 : 1.0	Actor
dvd = True	1 : 0	7.1 : 1.0	
event = True	1 : 0	7.1 : 1.0	
bore = True	0 : 1	6.6 : 1.0	
commentari = True	0 : 1	6.3 : 1.0	
flat = True	0 : 1	6.3 : 1.0	
knife = True	1 : 0	6.2 : 1.0	
rare = True	1 : 0	6.2 : 1.0	
told = True	1 : 0	6.2 : 1.0	
emot = True	1 : 0	6.0 : 1.0	
aw = True	0 : 1	5.8 : 1.0	
base = True	1 : 0	5.4 : 1.0	
melissa = True	0 : 1	5.4 : 1.0	Actor
john = True	1 : 0	5.2 : 1.0	Actor
perfectli = True	1 : 0	5.2 : 1.0	
worth = True	1 : 0	4.9 : 1.0	
african = True	1 : 0	4.9 : 1.0	
light = True	1 : 0	4.9 : 1.0	
taken = True	1 : 0	4.9 : 1.0	

TABLE X. Accuracy of the top 3 classifiers in categorizing the reviews based on the recommendation.

Reviewer Focus	Featureset	NB	MNB	BernoulliNB
Recommendation	raw-words	93.1	93.1	94.0
Recommendation	word-stem	92.2	92.2	93.1
Recommendation	words-nostop	91.4	91.4	92.2

H6 posited that viewers are attuned to the level of dispassion or extreme expression/passion in a movie review. Below is the accuracy of the top three classifiers in categorizing the reviews according to the Extreme-Emotion feature. From the results we can see that **H6** is weakly supported. Extreme subjective words such as “garbage" and “stupid" were used by the classifier to identify the category. Interestingly, though unfortunate, positive words such as “fabulous" and “intelligent" appeared less frequently than their negative counterparts in dispassionate reviews.

TABLE XI. Extremity: Most Informative Features

Word	Y : N	Ratio	Significance
garbag = True	1 : 0	15.7 : 1.0	Subjective
wow = True	1 : 0	10.0 : 1.0	
color = True	1 : 0	9.0 : 1.0	
suck = True	1 : 0	9.0 : 1.0	Subjective
stupid = True	1 : 0	8.8 : 1.0	Subjective
oh = True	1 : 0	8.1 : 1.0	
nice = True	0 : 1	7.7 : 1.0	
superhero = True	0 : 1	7.2 : 1.0	
crap = True	1 : 0	6.6 : 1.0	Subjective
sjw = True	1 : 0	6.2 : 1.0	Prejorative
space = True	1 : 0	6.2 : 1.0	
money = True	1 : 0	6.1 : 1.0	
4 = True	0 : 1	5.8 : 1.0	
mostli = True	0 : 1	5.4 : 1.0	
blow = True	1 : 0	5.2 : 1.0	Subjective
fabul = True	1 : 0	5.2 : 1.0	Subjective
intellig = True	1 : 0	5.2 : 1.0	Subjective
nobodi = True	1 : 0	5.2 : 1.0	
singl = True	1 : 0	5.2 : 1.0	
swear = True	1 : 0	5.2 : 1.0	Extremity
walk = True	1 : 0	5.2 : 1.0	
brilliant = True	1 : 0	4.9 : 1.0	Subjective
ladi = True	1 : 0	4.9 : 1.0	
hilari = True	1 : 0	4.6 : 1.0	Subjective
hear = True	0 : 1	4.4 : 1.0	

TABLE XII. Accuracy of the top 3 classifiers in categorizing the reviews based on the extremity.

Reviewer Focus	Featureset	NB	MNB	BernoulliNB
Extreme Emotions	raw-words	72.2	70.8	72.2
Extreme Emotions	word-stem	80.6	69.4	77.8
Extreme Emotions	words-nostop	72.2	70.8	70.8

LIMITATIONS

According to [2] machine learning is tuned through the training corpus (in the present case, our 1800 row database on 12 movies), and is vulnerable to over-training. In this case, one has to take risk of overfitting the model using limited size of data set.

This is a relatively small size of data set mainly due to the cost of creating a very specialized one, but is one of the unique contributions of the study. In [2], there was an warning that classification using the bag of words approach, based on raw words or word stems is susceptible to errors. The semantics of a word can change in different contexts.

DISCUSSIONS

Using Bernoulli, Naïve Bayes, and MNB classifiers, trained using a bag-of-words approach, we are able to provide evidence of agreement between classifiers and audiences, represented by persons who tagged reviews.

Audiences are not a homogeneous group. Some are emotive and others are more logical. Hence, some audience members care more about how a movie makes them feel the emotional experience than the technical aspects or production values of the movie. We show that movie viewers can relate to the emotionality expressed in a review in contrast to how technical the reviewer was. Therefore, some viewers will be influenced more by one than the other and extract more value from the review based on their own leanings.

Some folks take a more humanistic view over a factual approach in their world view. Hence, they focus more on the human expressions and interactions, growth and personality change of the hero over the span of the movie. The term “suspension of belief" refers to the viewer’s inability to accept the events in a movie, such as action sequences performed by disregarding the laws of physics. Unless everything could happen in real life, they cannot enjoy the movie. This difference in personalities determines what stands out in a movie with different viewers – also what resonates with them long after the movie ends.

Satisfaction is relative since we can be more or less satisfied watching a movie than its sequel or prequel. In a franchise such as Marvel or a series such as Aliens, there are natural comparisons to be made. But there are less reviewers who compare the sequel to the movie, than reviewers who simply review the movie on its own merit. This is important because previous movies in a series, tend to set certain expectations, making it hard to replace actors, or match CGI previously used, not to mention deliver a plot that is better than the previous, or stays true to the “Canon" already established in the series. Whereas for someone who is watching the movie without prior knowledge of the franchise, these differences are not even noticeable. However, variances can be problematic to its cult following who may obsess over them instead of enjoying the movie on its own merits.

Plot summaries are important in serving multiple purposes, which is why movie critics typically start by giving the major plot points, talking about the actors and discussing the theme, before launching in to their likes and dislikes. Summaries provide a framework within which to understand and relate to what the critic is espousing. Similarly, writing a paper review which starts with a summary, assures editors and authors that “yes, I read it properly and I do understand what it is trying to achieve." A Summary is a response to the rhetorical “Did we watch the same movie?". They ensure that the reviewer has not missed out on any major aspects of the movie. In other words, did the reviewer “get it"? Plot summaries also serve the purpose of informing viewers about what the movie entails, so the viewer can quickly surmise whether it is a story that would be interesting to them. For example, a viewer who just wants to be “transported to a nether realm" by a fantasy film may be more interested in special effects and have less interest in the plot points than someone wishing to watch a drama.

Whether a specific recommendation is given at the end of the review, is a window into the mindset of the reviewer. Some reviewers simply desire to express how they feel and what they enjoyed about the movie. Others specifically guide the viewer towards or away from the movie. Explicit recommendations are double-edged swords. On one hand, the reviewer demonstrates that they stand by what they said, bring their review to its arguably logical conclusion —- that of helping the viewer decide whether to spend time watching it or not. On the other hand, the reviewer could be engaged in reverse logic – overreaching for points that support what they want the potential viewer to do. This can be based on their own biases. The reviewer may feel the need to impose their world views on others, by endorsing or bashing movies with similar or dissimilar world views. This has less to do with the quality of the movie, and more to do with getting others aligned with their religion/politics/social views.

As with all the other dimensions, expressions in the extreme have their pros and cons. A dispassionate review conveys a feeling of balance, that the reviewer is approaching the movie from a neutral perspective, which suggests a more thorough exploration of the movie’s components. This has similarities to political centrism, often indicative of a

broader, less biased world view and arguably more logical train of thought. However, the lack of passion is a turn off for some. Centrists can come off as non-committed to truth and more interested in just getting along, by “sitting on the fence." In the same way, a dispassionate reviewer's blandness can give the impression that mediocrity is acceptable, thereby failing to stimulate the reader's decision, and therefore rendering the process pointless. Viewers want to be entertained, learn core lessons of life, be engrossed and so on. A passionate review can appear more aligned with their needs: finding a movie which gives them the maximum return in their time.

CONCLUSIONS

It was found in [5] that reviews help in two types of risk mitigation. The first is monetary risk, wasting money on a movie that is unsatisfying. The other factor is the prospect of getting a better movie to watch. The typical risk in relation to movies is $ 5 and 2 hours which you could use for something else, especially watching a free movie with ads. While time is precious, your emotional state is important as well. This emotional outcome tends to be ignored. It was argued in [15], risks that are too great to be effectively evaluated, simply get ignored. Such is the case with emotional risks of watching a movie which leaves the viewer upset. Without the use of Natural Language Processing, the review landscape becomes an emotional minefield. Step on the wrong review and your interest in the film gets blown up, causing you to miss a movie that could have been just apt for you. Similarly, with the wrong review, you get drawn into a messy story that leaves you unentertained, discouraged or filled with negative ideas – an ill-effect which can last for days. Without the use of NLP to help guide users, mismatched users/movies could not only reduce movie industry revenues and viewer entertainment, but also have wider negative impacts on the broader society. In this paper, we have sought to add to the usefulness of NLP in moving past the number of stars that a movie receives after being rated.

In the same way that a person seeking empathy will be more interested in a listening ear than a logical tongue, we believe that movie reviews can be made more relevant and effective by providing reviews that are aligned with the potential viewer's sense of what make a movie good. We hope to add this one more strategy to the repertoire of movie recommendation systems by pinpointing the movie reviews which will be more impactful towards the potential viewers' movie-watching experience.

ACKNOWLEDGMENTS

Authors acknowledge support from the US National Science Foundation under grant No. 1719514.

REFERENCES

1. E. Dichter, Harv. Bus. Rev. **44**, 147–166 (1966).
2. P. Chaovalit and L. Zhou, “Movie review mining: a comparison between supervised and unsupervised classification approaches,” (Proceedings of the 38th annual Hawaii international conference on system sciences, 2005) doi: 10.1109/HICSS.2005.445.
3. S. Sun, C. Luo, and J. Chen, Inf. Fusion **36**, 10–25 (2017).
4. K. Dave, S. Lawrence, and D. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” (WWW '03: Proceedings of the 12th international conference on World Wide Web, 2003) https://doi.org/10.1145/775152.775226.
5. D. Kahneman and A. Tversky, “Prospect theory: An analysis of decision under risk. in handbook of the fundamentals of financial decision making,” (World Scientific, 2013) pp. 99–127.
6. K. Topal and G. Ozsoyoglu, “Movie review analysis: Emotion analysis of imdb movie reviews,” (2016) doi:10.1109/ASONAM.2016.7752387.
7. E. Cambria and B. White, IEEE Comput. Intell. Mag. **9**, 48–57 (2014).
8. H. G. de Zúñiga, N. Jung, and S. Valenzuela, J. Comput. Mediat. Commun. **17**, 319–336 (2012).
9. M. W. Hughey and J. Daniels, Media Cult. Soc. **35**, 332–347 (2013).
10. S. I. Bhuiyan, Middle East Media Educator **1**, 14–20 (2011).
11. E. Gilbert and K. Karahalios, “Predicting tie strength with social media,” CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (ACM, 2009) pp. 211–220, doi: https://doi.org/10.1145/1518701.1518736.
12. W. Duan, B. Gu, and A. Whinston, Decis. Support Syst. **45**, 1007–1016 (2008).
13. C. M. Cheung and D. Thadani, “The effectiveness of electronic word-of-mouth communication: A literature analysis,” (2010) 23rd Bled eConference - eTrust: Implications for the Individual, Enterprises and Society.
14. J. Berger and K. L. Milkman, J. Mark. Res. **49**, 192–205 (2012).
15. S. Sen and D. Lerman, J. Interact. Mark. **21**, 76–94 (2007).

16. D. Sundaram, K. Mitra, and C. Webster, "Word-of-mouth communications: a motivational analysis," in *NA - Advances in Consumer Research*, Vol. 25, edited by J. W. Alba and J. W. Hutchinson (Association for Consumer Research, Provo, UT, 1965) pp. 527–531.
17. Y. Rao, Q. Li, L. Wenyin, Q. Wu, and X. Quan, Neural Netw. **58**, 29–37 (2014).
18. Y. C. Chang, C. C. Chen, Y. L. Hsieh, C. C. Chen, and W. L. Hsu, "Linguistic template extraction for recognizing reader-emotion and emotional resonance writing assistance," (ACL, 2015).
19. P. Parks, Media Cult. Soc. **41**, 1228–1244 (2019).
20. H. J. Gans, *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek, and Time (Medill Visions Of The American Press)* (Northwestern University Press, 1979).
21. S. Sood and X. Dreze, J. Consum. Res. **33**, 352–360 (2006).
22. Q. Hoang, "Predicting movie genres based on plot summaries," (2018), arXiv preprint arXiv:1801.04813.
23. A. Khan, M. A. Gul, I. Uddin, S. A. A. Shah, S. Ahmad, M. D. A. Firdausi, and M. Zaindin, Sci. Program. **vol. 2020** (2020).
24. G. Packard and J. Berger, J. Mark. Res. **54**, 572–588 (2017).
25. J. Mackiewicz, D. Yeats, and T. Thornton, IEEE Trans. Prof. Commun. **59**, 71–88 (2016).
26. J. Kim and P. Gupta, J. Bus. Res. **65**, 985–992 (2017).
27. B. Issac and W. Japp, "Implementing spam detection using bayesian and porter stemmer keyword stripping approaches," (TENCON 2009 - 2009 IEEE Region 10 Conference, 2009) pp. 1 – 5, doi: 10.1109/TENCON.2009.5396056.
28. H. Zhang, "The optimality of naive bayes," (Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004, 2004).