Determining the emotional experience evoked by films from online reviews

[Extended Abstract]

Osnat Mokryn*
Information Systems, University of Haifa
Haifa, Israel

David Bodoff
Nadim Bader
Yael Albo
Business Administration, University of Haifa
Israel

Joel Lanir
Information Systems, University of Haifa
Haifa, Israel

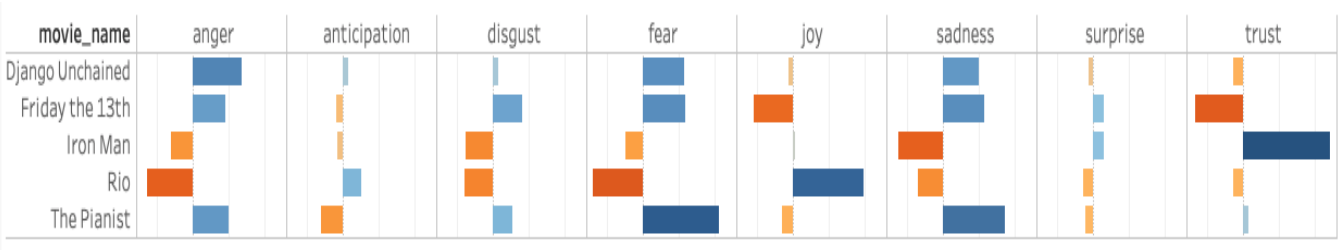

Figure 1: Sample emotional signature of five different films visualized with respect to the average value of each emotion

ABSTRACT

It is well established that reviews contain expressions of sentiment *towards* the reviewed items, e.g., "I liked it." Here, we hypothesize that in the case of experience goods and specifically films, the reviews also contain a signal of the emotions *evoked by watching* the movie, emotions that were experienced when watching it. That is, online reviews for experience goods also reflect the reviewer's emotions while experiencing the item, in the form of social sharing and can be reliably extracted from it. We postulate that the aggregated extracted emotional experiences for a movie form an emotional signature that reflects the emotions evoked by the movie. To establish that, we systematically conduct a set of analyses, each designed to offer evidence that supports our hypothesis. The ability to reliably, efficiently, and unobtrusively obtain the emotions evoked by films or other experience goods has numerous practical applications for both consumers and producers. For example, affective recommender systems can incorporate the film's evoked emotions as a feature. In addition, users' ability to easily see in advance the emotions that the item has induced in others, is relevant for research on the effect of expected emotions on decision-making.

KEYWORDS

emotions, experience goods, movies

1 INTRODUCTION

With the advancements in Internet technologies over the past two decades, communication platforms emerged that enable the sharing of information, experiences, and opinions in the form of online reviews [16]. Online reviews for products and services have surged in popularity, becoming a trusted and influential factor in consumers' decision process [9, 10]. Consumers consult reviews to obtain information about products' attributes and the experience of other consumers with them. When it comes to experience goods, i.e., products that can only be evaluated after consumption (e.g., books, music, hotels, restaurants and films), online consumer reviews have particular importance because the item can only be evaluated based on past users' experiences [21, 44].

The common wisdom in the computational social sciences is that online reviews reflect the consumers' opinions about the goods [3, 19]. Hence, the valence or the polarity of the reviews' sentiment became the subject of much research and interest [20, 30].

Here, we claim that reviews reflect the emotions experienced by the reviewer while consuming the item. that is, the emotions that are expressed in a review reflect not only what the person feels *towards* the movie, e.g., "I loved this film", but may also constitute a record of the emotions the person experienced while watching the film.

*An extended abstract of [27]

†Corresponding author: ossimo@gmail.com

WISDOM '20, held in conjunction with KDD'20, August 24, 2020, San Diego, CA USA, 2020

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In literary and philosophy studies, emotions are generally defined as having a target, or "intentionality". we distinguish between two cases: the emotion can be targeted at the movie as an artifact (e.g. admiration of the director or surprise at their choices, disappointment of an actor's performance, anger at a politically insensitive casting decision, etc.). This is what we call "emotions towards the movie". Or the emotions can be targeted at the (fictional or non-fictional) reality that is being depicted in the film or book, e.g. trust in the hero, or fear of the monster, or grief in response to a character's death. This is what we call emotions "evoked by watching" and what we aim to measure. Re-stated in these terms, the central claim of our work is that movie reviews will provide a signal, not only about viewers' emotions towards the artifact or the people who created it, but also towards *the characters and the situations being depicted.*

Our hypothesis – that reviews reflect emotions evoked by the item – is especially relevant in the case of an experience good, such as a movie. The main pillar in the experience of watching films is the emotional one. Indeed, the emotional experience is a major incentive for watching films [38]. The importance of emotions in films has long been recognized, as was the demonstrated ability of movies to elicit emotions in their audience. Affect elicitation, triggered by emotions, was found to be a powerful reason for box office success, and filmmakers use a variety of methods to elicit emotions in their audiences [11, 40]. Moreover, research in psychology shows that people share the emotions they'd experienced, and that sharing an emotional experience is an integral part of the experience [4, 7, 34].

Emotional experiences make us want to share them with others [4, 34]. People are eager to share their thoughts, experiences, and emotions publicly in front of friends and even strangers [41]. One form of social sharing is that of online reviews. Since a film is an emotional experience, and people share their emotional experiences, the following questions arise: (1) Do online reviews reflect people's emotional experience – the emotions that the movie evoked? And, (2) is it possible to automatically detect this signal? We hypothesize an affirmative answer to these questions. We emphasize that prior work has already established that reviews contain emotions, especially sentiment, *towards* the item as an artifact. Our analyses are designed to validate whether reviews also reliably reflect emotions that were experienced, i.e. evoked by watching the movie.

We employ a method that is similar to the one taken in [42] and [1]. It takes a bag-of-words approach and utilizes keywords in conjunction with a lexicon of emotions words to construct a per-movie emotional vector. We then proceed to establish our hypothesis that reviews include a signal of the emotions that people experienced while watching the film, and that our vectors capture this signal.

To establish that, we systematically conduct a set of analyses. We show that a film's emotional signature, calculated in the manner we propose, (a) makes intuitive sense; (b) is similar between sequels; (c) correlates with the emotions that subjects in an MTurk (Mechanical Turk) experiment report experiencing; (d) is similar to other movies of similar genre; (e) can predict what genre a movie belongs to and (f) can predict two film success indicators - rating and gross.

These findings support the central hypothesis that reviews contain a signal of the emotions experienced when watching the movie. Some of these pieces of evidence, especially 'c' (correlation with MTurk), are explicit and direct evidence of the hypothesis. Others are indirect yet powerful, especially when taken together, because the simplest explanation for them is that the emotional vectors indeed reflect the emotions evoked by watching the film. For example, it is difficult to see why emotions *towards* a movie would be genre-specific, while it is obvious (relevant literature will be cited below) that emotions experienced during the movie will depend on the genre. These and other analyses, supported by visualization tools, establish the meaning and potential usefulness of the measure.

Our work has important outcomes. First, The ability to reliably, efficiently, and unobtrusively obtain the emotions evoked by films or other experience goods has numerous practical applications for both consumers and producers. For example, affective recommender systems can incorporate the film's evoked emotions as a feature. The emotional affect the item evokes, when taken as a feature, has been shown to enrich recommendations [37]. Emotion-labeling of movies is normally a tedious affair [29], which our approach alleviates. Accessible visualizations of the emotional signature of films and other goods can enhance the consumer decision process. On the producer side, the ability to confirm what emotions a film evokes, and possibly relate that to success within different consumer groups, can be a novel windfall [11]. Second, the computational field of sentiment and emotion detection has until now simply assumed that emotions and sentiment-bearing words reveal the reviewer's sentiment about the movie. If the emotions words also reflect something else, as we suggest and intend to demonstrate, then any attempt to compute sentiment towards the item should attempt to separate the two.

Third, the problem is interesting because there is currently no known computational method that captures the emotions elicited by a film. A film is created with an intended emotional effect, and its success relies on the audience feeling these emotions: "A film is an invitation to feel in a particular way; however, while the audience can recognize how a film is cueing them to feel, they may either accept it or reject the invitation by not feeling those emotions" [38]. Indeed, while developing an emotional stimuli system that utilized film scenes, [11] found that it was hard to achieve a consensus as to the evoked emotions of even a single specific scene: "Examining the mean emotion ratings, we were struck by the variability among these 78 film [scenes] stimuli.. a scene in which a child falls and is rushed to a hospital room by his father.. [in] Kramer vs. Kramer produced levels of sadness that were not much greater than those for fear and surprise". So, nothing about this task is obvious. It is not obvious what emotions people will experience in a given film; it is not obvious that they will express those emotions in their reviews; and it is not obvious that such expression can be automatically detected. Our purpose is to establish all these in the affirmative.

2 BACKGROUND AND RELATED WORK

A large body of research has studied online reviews, with particular focus on the sentiment towards the item [20]. Here, we study whether reviews also reflect the consumer's emotional experience of the item, and whether that emotional experience can be reliably extracted.

2.1 Emotion as an attribute of experience goods

We adopt a conceptualization that considers the emotions that the item evokes in users. This conceptualization considers not what the item expresses or what the character experiences, but what the consumer experiences. The item is then characterized in terms of the emotions that it tends to evoke across individuals.

A motivating example stems from the psychological research on the basic emotion Surprise [7]. Surprise is a neutral (no valance) short-lived emotion that is the result of a misalignment between the expectations and the product/goods or its features (termed "schema discrepancy"). In that sense, Surprise is an emotion that arises from the experience with the product or goods. [7] found significant correlations between surprise and word-of-mouth. That is, Surprise about a product leads the consumer to want to share their experience. Their results also corroborate [34]'s findings on the active role of emotions in the need to share experiences.

An emotion is a complex chain of events that begins with a stimulus and includes feelings, psychological changes, impulses to act, and specific behavior. Emotions do not occur in isolation. Rather, they are responses to situations in an individual's life [36]. In his influential theory of emotions [31], Plutchik has arranged the basic emotions around a color-circle (commonly known as: "Plutchik's Wheel of Emotions")where the distance between two emotions around the circle reflects their similarity: adjacent emotions are more similar than emotions that are further removed, and contrasting emotions are placed opposite one another. According to his theory, two primary emotions could be combined to form a mixed or complex emotion. For example, Joy and Trust combine to create Love; Disgust and Sadness together create Remorse [32].

In this work, we follow Plutchik's theory of eight basic emotions to analyze and convey the emotion of films.

2.2 Identifying emotions in text

One popular approach for identifying emotions in text is a thesaurus/lexicon (knowledge)-based approach that contains synonyms and antonyms. The word-emotion association lexicon (NRC) [24], which we use here, compiled manual annotations for Plutchik's eight basic emotions. The lexicon was created by crowdsourcing to Mechanical Turk and contains over 14,000 words [24]. It was validated for several different domains, from fairy tales to discussions on the news, as well as U.S. sports fans' tweets during World Cup 2014 [18, 22, 23, 45]. For each word, it assigns a binary value for each of 8 emotions, based on whether the word is associated with that emotion. The majority of research on emotions-detection relies on a lexicon as a basic input, and adds additional computation to improve performance. In our case, we use a lexicon as-is. Our contribution is not in having developed a better algorithm for emotions prediction. Rather, our contribution is in systematically addressing the question posed at the outset, namely, whether the emotions-words that are used in film reviews (also) reflect the emotions evoked by watching the movie.

[28] create emotion vectors for each movie to serve as additional features in a collaborative filtering algorithm. They explore the importance of different types of features in providing good recommendations under different conditions. Our work differs in that rather than employing such vectors as machine learning features, we address the more basic question of whether a movie has an emotional signature that is reflected in its reviews. The work of [42] is more directly related to our work. Their focus, like ours, are the movies' emotional signatures themselves. They mostly employ a clustering approach, but the meaning of the results is not entirely obvious. For example, they report that some of the clusters are close to one another, and there is not a clear relationship between their clusters and genres. So, it is not clear what exactly their clusters tell us. What differentiates our work is primarily our systematic approach to establishing the validity of the measure. Each one of our analyses is designed to offer evidence that it is useful and meaningful to speak of a movie's emotional signature, and that our measure actually captures it.

3 CREATING EMOTIONAL SIGNATURES

For this study, we obtained a large dataset from the IMDb movie database site. IMDb offers a searchable dataset of over 185 million data items, from which we obtained over 1.5 million reviews and other metadata information for 9,666 films released by Hollywood between 1972 and July 2016.

To account for review volume and validity of temporal information, we consider only films released between 2003 and 2014. For example, it is hard to get the release weekend reviews for movies released before 1998, and recent movies may not have reached their full impact. We also do not consider films with less than 30 reviews in order to have a large enough sample for each film.

After data cleaning, our working dataset contained 2937 films with *at least* 30 reviews each, to a total of 717,498 reviews. Additional collected information contains the ratings, genres, cast, synopsis, budget, and box office revenue if it exists.

for calculating the *emotional signature* of each movie as extracted from its online reviews. Each review was annotated for Plutchik's eight basic emotions, similar to [1]. We used the NRC lexicon created by [25], which compiled manual annotations for the eight basic Plutchik's emotions as well as for positive and negative sentiment. The lexicon was created by crowd-sourcing to Mechanical Turk and has annotations for over 14,000 words. The prevalent approach to emotion detection in a text is based on the premise that the emotion expressed in the text is the aggregate of the emotions of the words comprising it. These techniques, therefore, look for the presence of appropriate affect words in the text. On top of being a very large database, it was proven to work very well on different domains [18, 22]. Each review was then annotated using the NRC lexicon for Plutchik eight emotions, as well as the positive and negative sentiment. Note that the lexicon may associate a given word with more than one emotion.

We created an emotional vector for each movie, termed its *emotional signature*, consisting of the eight emotions of the Plutchik's wheel. To achieve an emotional signature of a movie, we first aggregated all emotional words from all reviews in each of the eight Plutchik categories. Then, for each emotion, we count the number of occurrences of all words, i.e. from all reviews for that movie, that the lexicon associates with that emotion. This is the emotion-term-frequency vector.

The next step is normalization. Films differ in the number of reviews written for them, and reviews differ in length. To account for these differences, normalization may be in order. The emotional

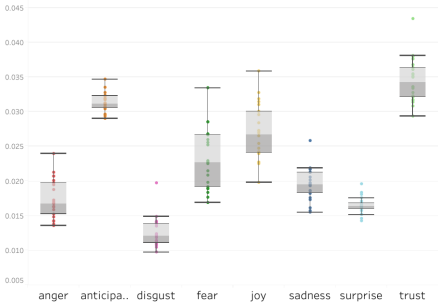

Figure 2: Distributions of relative emotion values across all movies showing large differences.

signature can be normalized by dividing each emotion by the total number of words in the reviews, or each emotion could be normalized according to its relative share of just the emotional words. However, the latter approach would not allow our analyses to account for the possibility that for some movies, the reviews include a higher percentage of words that express emotions, as compared with words that don't. Our modeling assumption is that this information is important, because movies that are more emotionally evocative may result in reviews that have a higher percentage of emotional text. Additionally, this type of normalization creates a direct dependency between the emotions, since the sum of their percentages would be forced to equal 100Therefore, to normalize, we calculated the number of words of each emotion out of the total amount of words that were written in all the reviews. This produced a vector of emotion values that represent the percentage or strength of each emotion in that movie.

In principle, this same approach can be carried out for various levels of granularity – at the level of a review, a film, or a genre. A *review document* is a single review. A *film document* refers to the aggregation of all reviews of the film as appear in the working dataset. A *genre document* refers to the aggregation of the reviews of all films tagged with this genre label, as appear in the working dataset. Formally, given a document $C \in \{Review, Film, Genre\}$ we define the emotional signature as a vector e of length eight, each entry corresponds to a basic emotion, as follows:

$$e_n = \frac{e_n^C}{M^C} \tag{1}$$

Where $e_n, n \in 1..8$ accounts for the eight emotions in the Plutchik wheel, e_n^C is the total number of emotion e_n words in C, and M^C is the total number of words in document C. In this work, we concentrate on the level of a film document. This is the appropriate level for our purpose and methods.

Figure 2 shows a boxplot of the distribution of emotions across all films in our dataset. The emotions have large variances, i.e. differences between films. Even this simple result is an initial indication that we are measuring *something*. It remains to offer (indirect) evidence that the signal represents what we assert, namely, the emotions that each movie evokes in viewers.

In the following sections, we continue to validate that the emotional signature indeed captures a film's evoked emotions. We design a set of validations, each designed to offer evidence that it is useful and meaningful to speak of a movie's emotional signature, and that our measure actually captures it.

4 AN EXPLORATION OF EMOTIONAL SIGNATURES IN FILMS

In this section, using mostly qualitative means, we explore the emotional signatures of movies. We start by visualizing the signature in two different manners. We continue by showing that sequels have significantly closer signatures than any arbitrary two films.

4.1 Visualizing Emotional Signatures of Films

Figure 1 depicts the emotional signatures of five selected films while emphasizing how those movies depart from the average by showing each emotion for each film compared to the mean value of that emotion across all films. It can be seen that the kids' animation film *Rio* has an emotional signature with very high Joy compared to the average and compared to the other movies, high Anticipation, and very low negative emotions such as Anger, Disgust, or Fear. On the other side of the scale is *The Pianist*, a movie about a Polish-Jewish musician's survival during world war II, which has the highest Fear and Sadness values among the five. Both *Django Unchained* and *Friday the 13th* are high on Anger, Sadness and Fear, but *Friday the 13th* is also high on Disgust, and very low on Joy compared to the average. *Iron Man* is very low on Sadness, and the highest on Trust, compared to the other four. These kinds of data explorations, which we found to have intuitive and reasonable interpretations, provide basic, if indirect, evidence that our measure captures what we intend, namely the emotions each movie evokes in viewers.

4.2 Emotional Signatures of Sequels

Our first approach to validating the instrument is to see if it behaves as we would expect in the case of sequels. Movie sequels are popular and profitable, but are hard to define [15], and are often classified in various ways [35]. One prominent factor in definitions of sequel, is the idea of Repetition and Re-experiencing. [15] relates to a sequel as a framework in which formulations of repetition are to be found. She points out that "Sequelisation as a form of repetition-compulsion is evidenced by the way sequels are designed to keep audiences coming back to cinema theatres, to re-experience the film". [13] asserts that "sequel production.. tend towards the formulaic, offering audiences more of the same".

Based on the above, if our emotions vectors really do measure what we claim, then sequels should exhibit mostly the same emotional signature as their base movie. This analysis represents a kind of validity check, sometimes called "face validity". Figure 3 depicts the emotional signatures of the *Harry Potter* sequels that are in our dataset,visualized using the Plutchik Radar [1]. We can see that the emotional signatures of the sequels in these examples show a strong level of similarity.

To test this hypothesis, we looked at all sequels in our database. We found Fifty sequels, ranging from 2 to 6 movies per sequel. We examined the distances between the emotional signatures. Given a base film M_i, $i \in [1, 50]$, and its follower in a sequel, M_i^f, the

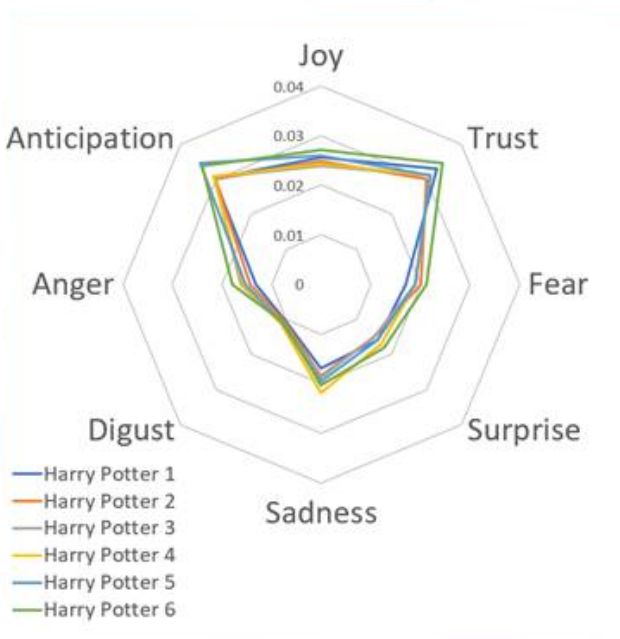

Figure 3: Emotional signatures of sequel movies: the six *Harry Potter* films in our dataset

Euclidean distance [8] between the sequel and the base is compared to the average distance of the sequel to other movies, and to its genre-movies.

For all movies results show that the sequel's emotions vector was significantly closer to the base film than to other films, ($M = 0.007, SD = 0.004$) distance from the base versus ($M = 0.019, SD = 0.005$) distance from other films. The result is highly significant ($t(49) = 16.5, p < .001$).

Next, we continue to compare the distance of the emotional signatures of the sequel to other films from the same genre. For reasons discussed below in section 6 this was perhaps most meaningful for 7 of the 50 cases, where the sequel and its base movie shared a single genre label. In addition, a few cases had to be dropped from the analysis because there were no other movies in the dataset that shared the same genre label combination as the sequel and its base. But whether considering all 46 sequels that could be analyzed, or only the 7 that had a single label, results are highly significant. For all 46 sequels that had other films (besides the base) in the same genre, $t(45) = 7.84, p < .001$.

Table 1 shows as an example the results for the seven movie pairs that had the same single genre label. In summary, sequels' emotions vector are much closer to their base film than to other films in the same genre. All together, results show that emotion signatures of sequels are significantly closer to their base than to other movies, even in the same genre. These results provide one piece of evidence that the measure, which is based on emotion-words in reviews, reflects the emotional experience.

Table 1: Sequels emotional signatures distance from each other compared to from their genre

Film Name	Genre	Distance from base film	Distance from Genre
Final Destination 5	Horror	0.0081	0.0157
Grown Ups 2	Comedy	0.0031	0.0117
Halloween II	Horror	0.0086	0.0090
Paranormal Activity 2	Horror	0.0103	0.0119
Scary Movie 4	Comedy	0.0106	0.0103
The Hangover Part II	Comedy	0.0029	0.0054
The Hills Have Eyes II	Horror	0.012	0.0159

5 CORRELATION WITH MANUAL MEASURES OF EVOKED EMOTIONS

For the next step in establishing that our measure reflects what we intend – namely, the emotions that watching a movie evokes – we compared the emotional signature of fifteen selected films with the explicit statements provided by survey participants regarding the emotions they experienced when viewing those films. This analysis demonstrates "convergent validity", a significant step in establishing that a proposed measure captures what is intended [5]. We ran our experiment using workers on Amazon's Mechanical Turk to maintain high external validity (i.e., provide us with a diverse pool of participants for the experiment).

We selected fifteen films from different genres. We included only films that had at least 100 reviews and were relatively well known (measured according to their total gross amount). We tried to get a wide range of genres, including at least two movies of each of the main genres of Action (and Adventure), Drama, Comedy, Animation, and Horror. Table 2 describes the films chosen for the experiment.

The experiment was set up as a survey on Amazon's Mechanical Turk (MT), which workers were asked to answer. Films were divided into three groups, and workers were asked to choose a group for which they have recently seen one or more of the films. After choosing a group, the worker was asked to choose the movie they were most familiar with. In order to validate that the worker has watched the selected movie, we then asked the worker to answer two questions related to the movie. For example, for the movie *Brides maids*, one of the questions asked: "The girls became sick from: (a) food poising; (b) eating too much fish; (c) the flu". Only workers who completed the two questions correctly were allowed to continue. Upon successful validation of the two questions, the worker was given an explanation regarding the emotions Finally, the worker was asked to rank each of the emotions for the movie he or she selected using a 7-point Likert-scale from very low to very high. A total of 527 workers accepted our HITs ("Human Intelligence Tasks"). Workers were able to accept the HIT three times each. The age range of our workers spanned from 16 to 66 years, and the mean age was 33.1 (SD: 10.5). The race distribution was: 41.7% Caucasian, 37.5% South Asian, 4.2% African, 4.2% East Asian, 1.1% Hispanic, and 11.3% other or unreported. We did not restrict participation

Table 2: Films chosen for the user evaluation and their characteristics as obtained from IMDb

Film Name	Genres	Reviews	M-Turk Respondents
Man of Steel	Action, Adventure, Fantasy	2429	31
Pirates of the Caribbean: The Curse of the Black Pearl	Action, Adventure, Fantasy	2096	31
The Avengers	Action, Adventure, Sci-Fi	1364	32
Mystic River	Crime, Drama, Mystery	901	35
The Conjuring	Horror, Mystery, Thriller	763	39
Harry Potter and the Deathly Hallows: Part 2	Adventure, Drama, Fantasy	733	43
Toy Story 3	Animation, Adventure, Comedy	699	40
The King's Speech	Biography, Drama, History	604	35
Mamma Mia!	Comedy, Musical, Romance	542	35
Bridesmaids	Comedy, Romance	469	36
The Help	Drama	445	37
Grown Ups	Comedy	305	35
The Devil Inside	Horror	279	34
Rio	Animation, Adventure, Comedy	162	34
What Happens in Vegas	Comedy, Romance	145	31

based on any demographics in order to have a fair sample of the Mechanical Turk (MT) worker population.

5.1 Evaluation of Results

For each film, we averaged the results of the MT survey data over the workers responding to questions for that film. Each film was hence represented by a vector of eight values ranging from 1 to 7, depicting the corresponding values of emotions given by the workers. This yielded 8 numbers for each of 15 films, to which we could compare the emotions-vectors that we computed from reviews.

We analyzed the data in four ways. First, we arranged the data into two long columns, with each row giving the emotions values for a single movie-emotion pair, according to both methods. The overall correlation between the two sets of numbers was 0.67 ($p < 0.05$).

Second, we anlayzed each movie separately. Results are presented in Table 3. For 11 of the 15 movies, there was a significant correlation (with only 8 numbers per correlation, statistical significance is not easily obtained). We interpret this result as adding evidence that our vector captures something closely related to the emotions that a movie evokes. In terms of what might cause some movies to have a better correlation than others, it could be that the MTurk data is noisier for some movies than for others. Indeed, we found that there is a correlation between how consistent the MTurk reports were for a given movie, and the degree of correlation with our emotions-vectors as was reported in Table 3. But this association was not statistically significant (corr = 0.4; n.s.). It is also possible that for some movies the computed vectors are less accurate. At this point we do not know whether some movies have an attribute that causes its reviews to be a less reliable representation of the experienced emotions.

Third, reversing the orientation, we examined which emotions showed higher correlation between the two methods, across the 15 films. Results are presented in Table 4. Fear was the emotion with

Table 3: Pearson correlations between computed vectors and MTurk results

Film Name	Pearson Correlation	Significance (p-value)
The King's Speech	0.864	p<.01
Mamma Mia!	0.863	p<.01
What Happens in Vegas	0.863	p<.01
Rio	0.857	p<.01
The Avengers	0.826	p<.05
Grown Ups	0.813	p<.05
Toy Story 3	0.793	p<.05
Bridesmaids	0.786	p<.05
Man of Steel	0.752	p<.05
Pirates of the Caribbean: The Curse of the Black Pearl	0.711	p<.05
Harry Potter and the Deathly Hallows: Part 2	0.710	p<.05
The Conjuring	0.702	p<.1
The Devil Inside	0.362	n.s.
The Help	-0.019	n.s.
Mystic River	-0.225	n.s.

the highest correlation over all films in the experiment set. Joy, Disgust and Sadness exhibited good correlations. Anger, Surprise and Trust, correlated moderately, yet were not significant. Anticipation correlated insignificantly and negatively. We suspect that a reason for this might be that the definition of Anticipation, and to a lesser extent Trust, and their relation to the movies was not clear enough to participants in the survey.

6 FILM GENRE LABELS

We next proceed to a number of analyses showing that our emotions-vectors characterize genres, a result that can be explained if the

Table 4: Spearman's rank correlation test over the different emotions

Emotions	Spearman's Rank Correlation	Significance
Fear	0.875	p<.01
Joy	0.699	p<.01
Disgust	0.609	p<.05
Sadness	0.588	p<.05
Anger	0.389	n.s.
Surprise	0.366	n.s.
Trust	0.239	n.s.
Anticipation	-0.153	n.s.

emotions vectors reflect the emotions that were evoked by watching the film. A "genre film" serves to indicate a form of seriality that "through repetition and variation, tell[s] familiar stories with familiar characters in familiar situations" [17]. Moreover, the genre of any films fulfills a simple descriptive and classificatory function that aims to situate and identify a film [26]. Strong correlations, sometimes surprising, are known to exist between genres and emotions. For example, "audiences are attracted to Horror and Drama movies even though negative and ambivalent emotions are likely to be experienced [2]. Because of the known association between genre and emotion, genre is an additional perspective for analysis, to further explore the validity of our emotions-vectors.

Our dataset consists of 2, 937 films, and a total of 21 different genre labels. We expected that most films would be assigned a single genre label. If individual genres also have a typical emotions profile, then genres would be a good perspective for analyzing our vectors. But we found that of the 2,937 films, only 309 had been assigned a single label. Furthermore, those films with a single label were mostly concentrated into three genres – Horror only (58), Comedy only (107), and Drama only (124). Many other genres had no films whatsoever assigned only that genre, or very few: Only Action: 4; War: 0; Sport: 0; Western: 0; Animation: 0; Biography: 0; Adventure: 3; Crime: 0; Drama: 124; Family: 0; Fantasy: 1; Musical: 0; Music: 0; Romance: 0; Sci-fi: 0; Thriller: 8; Mystery: 0; Documentary: 4; Horror: 58; History: 0; Comedy: 107.

Most films had been assigned 2,3, or 4 labels, as follows: One label: 309; Two labels: 806; Three labels: 977; Four labels: 584; Five labels: 207; Six labels: 37; Seven labels: 15; Eight labels: 2.

Our analyses using genres begin with films that were solely assigned to Horror (58 films), Comedy (107), or Drama (127). The reason is that we are using genre as a proxy to group films by the emotions they evoke, in order to test the validity of our emotions-vectors. For movies that have multiple genres, it is not obvious that they will belong reliably to a set of films that evoke a certain type of emotion. Moreover, the genre-combinations were surprisingly dispersed, with over 624 genre combinations, so there is almost no data in each such group. Table 5 shows the top snippet of the alphabetical list of genre-combinations.

As shown, almost no genre-combinations have more than a tiny number of films. This makes it difficult to conduct a quantitative analysis based on those combo-genres. This is also why in Section 4.2 above, we only compared to their genre-baseline those

Table 5: Example of # of films assigned the same Multi-genre labels

Genres	# films with same genre labels
Action	4
Action, Adventure	6
Action, Adventure, Biography, Crime, History, Romance, Western	1
Action, Adventure, Biography, Drama, History, Romance, War	1
Action, Adventure, Biography, Drama, History, War	1
Action, Adventure, Comedy	7
Action, Adventure, Comedy, Crime	3
Action, Adventure, Comedy, Crime, Family, Romance, Thriller	1
Action, Adventure, Comedy, Crime, Thriller	2
etc.	etc.

sequels that belonged solely to Horror, Comedy, or Drama; all other genre-combinations are both noisy and also have very few examples from which to construct a genre baseline. In the Discussion Section, we reflect on the usefulness of genre-labels, in light of the apparent need for so many idiosyncratic genre-label combinations. The genre-based analyses that follow are limited to those three genres and the films that are assigned solely to them.

6.1 Genre Average

Our first analysis calculates genre-averages for Drama, Horror, and Comedy, considering only films that were assigned only that single label. We compare them to see if they differ in a way that accords with what we would expect. This is a kind of "face validity" test. The three vectors are shown in Table 6. Two observations accord with what we would expect. First, within each genre, the relative strengths of different emotions make intuitive sense. Second, the total amount of emotion increases from Comedy, to Drama, then Horror. This, too, aligns with what we might expect.

In a second analysis, we analyzed whether the emotions-vectors were more alike within genres than between genres, as one would expect if the vectors reflect experienced emotions. For each of the three genres, we compared the average distance between two movies within the genre, and between one movie from that genre and a second movie from a different genre. Results are shown in Table 7. It can be seen that the emotions-vectors behave as expected. Taken together, these analyses provide additional evidence of "face validity" of the proposed way to measure evoked emotions.

Finally, we wanted to do a bigger analysis that includes all movies, even though most belong to genre-combinations with only a tiny number of other films. To overcome this problem, we define a measure of "genre similarity" between two movies. For each film, we generate a vector of 21 dummy variables representing its binary assignment w.r.t. each of the 21 different solo genres. We define the

Table 6: Genre Averages

Genre	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	Total
Comedy	.016	.032	.014	.019	.029	.017	.017	.033	.177
Drama	.018	.033	.013	.024	.028	.023	.017	.039	.195
Horror	..027	.034	.022	.036	.023	.029	.022	.030	.224

Table 7: Pairwise distance within and between genres

Genre	Average within-genre distance	Distance between one movie from this genre and one from another genre
Comedy	.010	.016
Drama	.013	.016
Horror	.012	.024

genre similarity between two films as the inverse of the Hamming distance between two movies on their 21-bit vectors.

We find that there is a highly significant correlation (corr=0.17, $p < .01$) between the Hamming distance in genres, and the Euclidean distance of emotions-vectors. That is, *the more different the two movies are in the genre-labels, the more different they are in their emotions-vectors*.

Taken together, these analyses provide clear support for our hypothesis that the reviews' emotions-words at least partly represent the emotions evoked by the movie.

6.2 Predicting Genres According to a Film's Emotions

Given the apparent distinctiveness of the emotions-vectors for the three genres, it may be possible to identify a movie's genre based on its emotional signature. If this prediction succeeds, it will provide another form of evidence – sometimes called "criterion-related validity" – that emotional signatures are meaningfully conceived and measured. Criterion-related validity means that the construct, when measured as one proposes, predicts something else that one would expect, if the measure means what is claimed. The purpose is not to build the best predictive model of the target. Rather, the purpose is to add one more type of evidence that the proposed measure is indeed capturing what is intended, at least partly. In this particular case, the evidence is particularly powerful, because we use the proposed emotional signature to predict genre, an attribute of the film *that is known to be related to emotions*. We view this, if successful, as a particularly strong type of criterion-related validity.

Our task is to determine the genre of a film from its emotional signature. The emotional signature is a vector of length eight, with each entry corresponding to a basic emotion, and its value represents its strength in the aggregated text of the reviews. We trained several classifiers as well as an ensemble of classifiers [33] utilizing WEKA [12] with its default parameters. Specifically, we used Logistic regression, J48 decision trees, SVM, Naive Bayes, K-nearest neighbor (k=7), random forest, and two ensemble methods, namely Bagging and Adaboost [14, 39]. As before, the test-set for this experiment consists of films with a single genre label, for the following three genres: Drama (128 films), Comedy (107 films), and Horror (57 films). We conduct 10-fold cross-validation experiments on the test-set.

Our experimental results (Table 8) show that the emotional signatures predict the genre with an average accuracy of 0.9 and an average AUC of 0.968. Hence, when a film belongs to one genre, it can be predicted by the emotional signature extracted from the online reviews. We note in passing that [42] found a less straightforward relationship, but they had studied the relationship between genres and signature-clusters, whereas we directly analyzed genres versus the signatures of individual films. Our finding lends additional evidence to the conceptualization and measurement of emotional signatures. Moreover, the result hints at a possible new basis for understanding, or possibly even defining genres in terms of the emotions that they evoke.

7 SUCCESS PREDICTION

We continue by exploring whether emotional signatures can predict film success. For success indicators, we take the rating of the film and its box office receipts. The first model will attempt to predict average movie rating across all reviewers, using emotions vectors as predictors. The second model will attempt to predict box office receipts, using the emotions vectors and average movie rating as predictors. We expect that the emotional experiences of viewing the film will be predictive of its success. Success in this task will represent one more piece of evidence – a classic kind of "criterion-related validity" – that emotional signature is a meaningfully conceived concept, and adequately measured by our methods.

Predicting film ratings [43] and box office gross receipts [6] from online reviews have been the subject of much research. Finding a "best" predictor is out of the scope of this paper. Rather, as stated, our purpose is to validate the relation between the affect elicitation that is triggered by emotions, as captured by the emotional signature, and success [11].

Our predictive models include not only the movie's emotional signature but the genre as well, and most importantly, the interaction between the two. The reason is that ratings and enjoyment may depend on what kind of experience one was expecting or intending when choosing the film. In the extreme example, sometimes a viewer seeks a "negative" emotion such as fear and will choose a Horror movie in order to experience that. In fact, a whole body of literature seeks to understand how it can be that a rational person would want to experience a negative emotion, as summarized, for example, in [2]. Whatever the reason, the effect of a given emotional experience may depend on what the movie-goer wanted; in terms of our variables, the model is that the effect of the emotions-vector

Table 8: Genre prediction

		Comedy		Drama		Horror		W. Average	
	Acc.	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Logistic Reg.	0.906	0.898	0.967	0.902	0.961	0.931	0.972	0.907	0.968
Decision Tree	0.837	0.834	0.89	0.835	0.859	0.837	0.911	0.837	0.881
SVM	0.816	0.81	0.854	0.793	0.821	0.881	0.907	0.817	0.805
Naive Bayes	0.778	0.763	0.914	0.737	0.894	0.778	0.993	0.778	0.921
KNN (k=7)	0.861	0.845	0.96	0.838	0.951	0.94	0.996	0.861	0.963
Random Forest	0.871	0.86	0.964	0.865	0.963	0.908	0.995	0.872	0.97
Bagging	0.871	0.866	0.953	0.861	0.938	906	0.98	0.872	0.952
AdaBoostM1	0.65	0.725	0.763	0.4	0.651	0.791	0.951	0.599	0.753

on success will depend on the genre. For example, sadness might be a good experience – one that causes high ratings – for a drama film, but not for a musical.

For reasons of autocorrelation, in this model we combined the four negative emotions (Anger, Disgust, Fear, Sadness) into a single variable, so that success is predicted by four positive emotions plus one combined negative emotion. We construct two separate predictive models, one to predict box office gross receipts, and the other to predict the average numeric rating that was given in the reviews.

Table 9: Results of model predicting movie rating from emotional signature

Variable	Parameter Estimate	Pr > F
Adventure	1.39	$p < 0.0001$
Music	3.97	$p < 0.0001$
Negative Emotions	-23.74	$p < 0.0001$
Trust	36.29	$p < 0.0001$
Surprise	138.94	$p < 0.0001$
Anticipation	-39.32	$p < 0.0001$
Adventure x Negative Emotions	-16.88	$p < 0.0001$
Drama x Negative Emotions	18.91	$p < 0.0001$
Family x Negative Emotions	-8.891	$p < 0.0001$
Biography x Negative Emotions	3.92	$p < 0.0001$
Music x Negative Emotions	-24.81	$p < 0.0001$
Documentary x Negative Emotions	30.14	$p < 0.0001$
Drama x Joy	40.04	$p < 0.0001$
Animation x Joy	20.88	$p < 0.0001$
Documentary x Joy	-39.68	$p < 0.0001$
Music x Trust	-77.53	$p < 0.0001$
Horror x Trust	-10.42	$p < 0.0001$
Drama x Surprise	-118.81	$p < 0.0001$
Comedy x Anticipation	-13.33	$p < 0.0001$
Romance x Anticipation	-6.04	$p < 0.0001$

Results for the model predicting movie rating are shown in Table 9. Two genre had significant main effects, as did four emotion. Many genre-emotion interactions were significant. Results for the model predicting box office receipts are shown in Table 10. Movie rating was significant, as expected. A number of genre main effects were significant. Only anticipation was significant as an emotion main effect. Many genre-emotion interactions were significant predictors.

Table 10: Results of the model predicting movie box office receipts from emotional signature

Variable	Parameter Estimate	Pr > F
Movie rating	0.16	$p < 0.0001$
Anticipation	13.49	$p < 0.0001$
Adventure x Negative Emotions	-13.46	$p < 0.0001$
Action x Negative Emotions	-11.5	$p < 0.0001$
Adventure x Joy	-44.49	$p < 0.0001$
Family x Joy	44.84	$p < 0.0001$
Adventure x Trust	-39.58	$p < 0.0001$
Drama x Trust	-7.04	$p < 0.0001$
Fantasy x Surprise	-35.63	$p < 0.0001$
Family x Anticipation	-38.45	$p < 0.0001$

In both models, the genre-emotion interactions are interesting, and are also the results that most obviously support our hypothesis. The pattern of these interaction results make sense if the emotion vectors reflect the emotions that people experienced during the film, not emotions towards the artifact. Our interpretation is that genre is acting as a surrogate for the kinds of emotions that the viewer was seeking or expecting.

8 DISCUSSION AND CONCLUSIONS

We have shown that emotions extracted from online reviews for experience goods are reflective of the experience with them, over the film domain. Our results were validated through a series of validations, each offering additional evidence.

There are several limitations to our work. The MTurkers had seen the movie but at various times before completing the survey, while most of the people in the IMDb group had probably written the reviews close to viewing the movie. On the other hand, if anything, this would undermine the attempt to find that our signatures correlate with the emotions that people say they experienced when asked. Also, neither the IMDb reviewers nor the MTurkers may reflect the entire population. For example, IMDb reviewers may be more opinionated. This, too, can be further evaluated.

Our work also raises new technical questions. In particular, if our claim is supported, it means that reviews' emotions-words reflect a mixture of two different things – their sentiment or emotions towards the movie artifact, and also the emotions they experienced while watching. This poses a new technical challenge, to separate the two. This is necessary, regardless of which signal one seeks. For example, there is much work that intends to extract emotions and sentiment towards the movie artifact. Those algorithms can be improved, if they are able to first extract that aspect alone. This remains an open challenge.

There are many future directions to this work, from the research of the method in the context of other experience goods, to the research of the implications in recommender systems or marketing. Other directions also exist. The question of how to visualize not a few, but thousands of items' emotional signatures is challenging. Another research direction is the understanding of genres and their emotional signatures. Can close-by genres be determined by films' emotional signatures? Can an emotional definition for a genre be then devised?

Lastly, success prediction and emotions can be further studied for additional experience goods such as restaurants or hotels.

REFERENCES

[1] Nadeem Bader, Osnat Mokryn, and Joel Lanir. 2017. Exploring emotions in online movie reviews for online browsing. In *Proceedings of the 22nd international conference on intelligent user interfaces companion*. ACM, 35–38.

[2] Anne Bartsch, Markus Appel, and Dennis Storch. 2010. Predicting emotions and meta-emotions at the movies: The role of the need for affect in audiences' experience of horror and drama. *Communication Research* 37, 2 (2010), 167–190.

[3] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems* (2013), 1541–1672.

[4] Véronique Christophe and Bernard Rimé. 1997. Exposure to the social sharing of emotion: Emotional impact, listener responses and secondary social sharing. *European Journal of Social Psychology* 27, 1 (1997), 37–54.

[5] Linda Crocker and James Algina. 1986. *Introduction to classical and modern test theory*. ERIC.

[6] Chrysanthos Dellarocas, Xiaoquan Michael Zhang, and Neveen F Awad. 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing* 21, 4 (2007), 23–45.

[7] Christian Derbaix and Joëlle Vanhamme. 2003. Inducing word-of-mouth by eliciting surprise – a pilot investigation. *Journal of Economic Psychology* 24, 1 (2003), 99 – 116. https://doi.org/10.1016/S0167-4870(02)00157-5

[8] Michel-Marie Deza and Elena Deza. 2006. *Dictionary of distances*. Elsevier.

[9] Wenjing Duan, Bin Gu, and Andrew B Whinston. 2008. Do online reviews matter?—An empirical investigation of panel data. *Decision support systems* 45, 4 (2008), 1007–1016.

[10] Kristopher Floyd, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling. 2014. How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing* 90, 2 (2014), 217–232.

[11] James J Gross and Robert W Levenson. 1995. Emotion elicitation using films. *Cognition & emotion* 9, 1 (1995), 87–108.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.

[13] Stuart Henderson. 2017. *The Hollywood sequel: history & form, 1911-2010*. Bloomsbury Publishing.

[14] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. Vol. 398. John Wiley & Sons.

[15] Carolyn Jess-Cooke. 2009. *Film Sequels: Theory and Practice from Hollywood to Bollywood: Theory and Practice from Hollywood to Bollywood*. Edinburgh University Press.

[16] Andreas M Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons* 53, 1 (2010), 59–68.

[17] GRANT Barry Keith. 1995. Film genre reader II. Austin.

[18] Alistair Kennedy, Anna Kazantseva, Diana Inkpen, and Stan Szpakowicz. 2012. Getting emotional about news summarization. In *Canadian Conference on Artificial Intelligence*. Springer, 121–132.

[19] Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 1363–1367.

[20] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5, 1 (2012), 1–167.

[21] Nikos Malandrakis, Alexandros Potamianos, Georgios Evangelopoulos, and Athanasia Zlatintsi. 2011. A supervised approach to movie emotion tracking. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2376–2379.

[22] Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop*. 105–114.

[23] Saif Mohammad. 2012. Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics*. 587–591.

[24] Saif M Mohammad and Peter D Turney. 2012. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* (2012).

[25] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[26] Raphaëlle Moine. 2009. *Cinema genre*. John Wiley & Sons.

[27] Osnat Mokryn, David Bodoff, Nadim Bader, Yael Albo, and Joel Lanir. 2020. Sharing emotions: determining films' evoked emotional experience from their online reviews. *Information Retrieval Journal* (2020). https://doi.org/10.1007/s10791-020-09373-1

[28] Yashar Moshfeghi, Benjamin Piwowarski, and Joemon M Jose. 2011. Handling data sparsity in collaborative filtering using emotion and semantic based features. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 625–634.

[29] Ante Odic, Marko Tkalcic, Jurij F Tasic, and Andrej Košir. 2012. Relevant context in a movie recommender system: Users' opinion vs. statistical detection. *ACM RecSys* 12 (2012).

[30] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2, 1-2 (2008), 1–135.

[31] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience* 1, 3 (1980), 3–33.

[32] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.

[33] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B Kantor. 2015. *Recommender systems handbook*. Springer.

[34] Bernard Rimé, Pierre Philippot, Stefano Boca, and Batja Mesquita. 1992. Long-lasting Cognitive and Social Consequences of Emotion: Social Sharing and Rumination. *European Review of Social Psychology* 3, 1 (1992), 225–258.

[35] Subhadip Roy and Shilpa Bagdare. 2017. Why Do We Watch Sequels? A Qualitative Exploration from India (Structured Abstract). In *Creating Marketing Magic and Innovative Future Marketing Trends*. Springer, 1443–1448.

[36] Klaus R Scherer et al. 1984. On the nature and function of emotion: A component process approach. *Approaches to emotion* 2293 (1984), 317.

[37] Yue Shi, Martha Larson, and Alan Hanjalic. 2010. Mining mood-specific movie similarity with matrix factorization for context-aware recommendation. In *Proceedings of the workshop on context-aware movie recommendation*. ACM, 34–40.

[38] Greg M Smith. 2003. *Film structure and the emotion system*. Cambridge University Press.

[39] Dr B Srinivasan and P Mekala. 2014. Mining Social Networking Data for Classification Using REPTree. *International Journal of Advance Research in Computer Science and Management Studies* 2, 10 (2014).

[40] Ed S Tan. 2013. *Emotion and the structure of narrative film: Film as an emotion machine*. Routledge.

[41] Sharon Y Tettegah and Dorothy L Espelage. 2015. *Emotions, technology, and behaviors*. Academic Press.

[42] Kamil Topal and Gultekin Ozsoyoglu. 2016. Movie review analysis: Emotion analysis of IMDb movie reviews. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 1170–1176.

[43] P. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*.

[44] Rahat Ullah, Naveen Amblee, Wonjoon Kim, and Hyunjong Lee. 2016. From valence to emotions: Exploring the distribution of emotions in online product reviews. *Decision Support Systems* 81 (2016), 41–53.

[45] Yang Yu and Xiao Wang. 2015. World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans' tweets. *Computers in Human Behavior* 48 (2015), 392–400.