

**Feeling is Believing:
Viewing Movies Through Emotional Arcs**

by

Eric Chu

B.S., University of California, Berkeley (2014)

Submitted to the Program in Media Arts and Sciences
in partial fulfillment of the requirements for the degree of

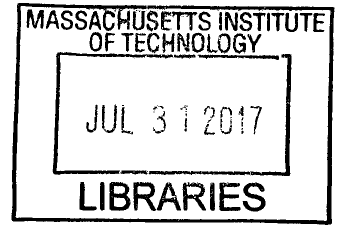
Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© Massachusetts Institute of Technology 2017. All rights reserved.



ARCHIVES


Signature redacted

Author

Program in Media Arts and Sciences
May 12, 2017

Signature redacted

Certified by

 Deb Roy
Associate Professor
Program in Media Arts and Sciences
Thesis Supervisor

Signature redacted

Accepted by

 Pattie Maes
Academic Head
Program in Media Arts and Sciences

**Feeling is Believing:
Viewing Movies Through Emotional Arcs**

by
Eric Chu

Submitted to the Program in Media Arts and Sciences
on May 12, 2017, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

This thesis uses machine learning methods to construct emotional arcs in movies, calculate families of arcs, and demonstrate the ability for certain arcs to predict audience engagement. The system is applied to Hollywood films and high quality shorts found on the web.

We begin by harnessing deep convolutional neural networks for audio and visual sentiment analysis. These models are trained on both new and existing large-scale datasets, after which they can be used to compute separate audio and visual emotional arcs for any video. We then crowdsource annotations for 30-second video clips extracted from highs and lows in the arcs in order to assess the micro-level precision of the system. Precision is measured in terms of agreement in polarity between the system's predictions and annotators' ratings. The final model combining audio and visual features achieves a precision of 0.894.

Next, we look at macro-level characterizations of movies by investigating whether there exist 'universal shapes' of emotional arcs. In particular, we develop a clustering approach to discover distinct classes of emotional arcs.

Finally, we show on a sample corpus of short web videos that certain emotional arcs are statistically significant predictors of the number of comments a video receives. These results suggest that the emotional arcs learned by our approach successfully represent macroscopic aspects of a video story that drive audience engagement. Such machine understanding could be used to predict audience reactions to video stories, ultimately improving our ability as storytellers to communicate with each other.

Thesis Supervisor: Deb Roy
Title: Associate Professor
Program in Media Arts and Sciences

**Feeling is Believing:
Viewing Movies Through Emotional Arcs**

by

Eric Chu

The following people served as readers for this thesis:

Signature redacted

Thesis Reader

A redacted signature consisting of a curved line with two small dots above it.

.....

Pattie Maes
Professor
MIT Media Lab

Signature redacted

Thesis Reader .

A small redacted signature mark.

.....

Hugo Larochelle
Research Scientist
Google

Acknowledgments

A sentence or less for each of the following people hardly seems fair. The brevity belies the huge appreciation I have for everyone here.

A first thank you goes to my advisor Deb Roy for guiding me in my research, giving me the freedom to explore, and encouraging me to work on this topic, which started off as just a fun side project. I would also like to thank my readers Hugo Larochelle and Pattie Maes for their valuable feedback, in addition to serving as inspiration as researchers.

Two years flies by when you're in good company. A big thank you goes to all the wonderful characters in LSM, a group held together by the indomitable Heather Pierce. A special thank you goes to my fellow grad students and colleagues – Ann Yuan, Anneli Hershman, Eric Pennington, Ivan Sysoev, Juliana Nazaré, Lisa Conn, Luke Wang, Martin Saveski, Mina Soltangheis, Nabeel Gillani, Nazmus Saquib, Neo Mohsenvand, Perng-hwa “Pau” Kung, Prashanth Vijayaraghavan, Raphael Schaad, Sneha Priscilla Makini, Sophie Chou, and Soroush Vosoughi – each of whom means much more to me than I let on.

Thank you also to Dominik Martinez and Sabrine Ahmed Iqbal, the UROPs whose work made my research possible.

A big shout out goes to my other Boston friends, both old and new. In particular, thank you to Michael Song, Charlotte Zhu, and Eddy Awad for the good times.

No acknowledgments would be complete without thanking my long-distance friends – the day ones (you know who you are), Michael Lebow, and Giacomo Cupido. Here's to a million more group texts and secret trips.

Most of all, thank you to my mom, dad, and brother for their continual love and support, without which none of this would be possible.

Contents

1	Introduction	17
1.1	Contributions	18
1.2	Overview	19
1.3	Outline	21
I	Modeling - feeling from sights and sounds	23
2	Constructing emotional arcs	25
2.1	Video datasets	25
2.2	Methodology	27
3	Image modeling	31
3.1	Sentibank dataset	31
3.2	Model	32
3.3	Sentiment prediction	33
3.3.1	Sentiment and color - image scrambling experiments	35
3.4	Emotional biconcept prediction	37
4	Audio modeling	41
4.1	Spotify dataset	41
4.2	Audio representation	43
4.3	Model	44
4.4	Sentiment prediction	45

4.5	Uncertainty estimates	46
5	Finding families of emotional arcs	49
5.1	Approach: clustering using k-medoids and dynamic time warping . . .	49
5.2	DTW formulation	50
5.3	LB-Keogh for speed-up <i>and</i> better modeling	51
5.4	Practical notes	53
5.5	Cluster results	54
II	Evaluation - moments and arcs	57
6	Crowdsourcing ground truth	59
6.1	Video clip extraction	60
6.2	Crowdsourcing experiment	63
6.2.1	Instructions and questions	63
6.2.1.1	Task	63
6.2.1.2	Instructions	63
6.2.1.3	Questions	65
6.2.1.4	Design considerations	65
6.2.2	Experimental setup and stats	66
6.2.2.1	Stats	66
6.2.2.2	Quality control	67
6.3	Definitions and terminology	69
6.4	Difficulty of task: inter-annotator agreement and confidence of ratings	70
6.4.1	Ambiguous clips	70
6.4.2	Variance of valence ratings	70
6.4.3	Confidence ratings	71
6.5	Accuracy of system	72
6.5.1	Defining precision	72
6.5.2	Overall precision	73
6.5.3	Precision by various cuts	74

6.5.4	Precision of audio	75
6.5.5	Precision by genre	75
6.6	Combined audio-visual model	77
6.6.1	Model and features	77
6.6.2	Movie embedding features	78
6.6.3	Final combined precision	80
6.6.4	Combined arc	80
7	Engagement analysis	83
7.1	The effect of emotional content features	84
7.1.1	Results	85
7.2	The effect of certain emotional arcs	85
7.2.1	Results	86
7.2.1.1	Good shape #1.	86
7.2.1.2	Good shape #2	90
III	Wrap up - where we are and beyond	95
8	Related work	97
8.1	Emotional arcs	97
8.2	Sentiment analysis, deep learning, and affect analysis	98
8.3	Computational approaches to understanding story	99
9	Limitations and future work	101
9.1	Limitations	101
9.2	Future work	102
9.2.1	Extensions	102
9.2.2	New Directions	106
10	Conclusion	109
10.1	Summary of work	109

10.2 Final remarks	110
A A second method for finding families of arcs	111
A.1 Formulation	111
A.2 Basis results	113
B Other crowdsourcing results	115
B.1 Q3: Emotions in clips	115
B.2 Q4: Audio, dialogue, visual in clips	117
C Combined audio-visual model example	119

List of Figures

1-1	Overview	20
2-1	Films Corpora: distribution of release year	26
2-2	Shorts Corpora: distribution of duration	27
2-3	Visual emotional arc	27
2-4	Emotional arcs: effect of smoothing values	28
2-5	Audio emotional arc	29
3-1	Example emotional biconcept	32
3-2	Image scrambling	35
3-3	Image scrambling: varying block sizes	36
3-4	Biconcept predictions	39
4-1	Spotify dataset: distribution of valence and energy features	42
4-2	Spotify dataset: relationship between valence and energy.	43
4-3	Example mel-spectrogram	44
5-1	Pathological example of how k-means and Euclidean distance fails for clustering emotional arcs	51
5-2	Warping windows for Keogh lower bound	52
5-3	Envelopes for Keogh Lower Bound	53
5-4	Elbow plots for k-medoid clustering	55
5-5	K-medoids on Films Corpora for k=5	56
5-6	K-medoids on Shorts Corpora for k=5	56

6-1	Peak detection example on the visual emotional arc of <i>Fantastic Mr. Fox</i>	62
6-2	Clip extraction edge case: overlapping peaks/valleys	62
6-3	Number of clips annotated per worker	66
6-4	Variance of valence ratings	70
6-5	Q2: Mean valence ratings, variance of valence ratings, and mean confidence ratings	71
6-6	Q2: Confidence ratings	72
6-7	Variance of valence vs. Confidence	73
6-8	Valence Ratings: Cuts	74
6-9	Creating movie embeddings using biconcept classifier	79
6-10	TSNE of movie embeddings, overlaid with genre	79
6-11	Combined emotional arc for the movie <i>Her</i>	81
7-1	Shorts Corpora: distribution of number of comments	83
7-2	Statistically significant arc in $k = 5$ clustering	87
7-3	Statistically significant arc in $k = 8$ clustering	90
7-4	Statistically significant arc in $k = 10$ clustering	93
A-1	Top modes for Films Corpora	112
A-2	Top modes for Shorts Corpora	112
A-3	Most similar films to mode 3-flipped	113
A-4	Most similar films to mode 6	114
B-1	Pairwise occurrences of emotions	116
C-1	Combined model: audio and visual arcs used for example	119

List of Tables

3.1	Modified Alexnet architecture used for image classification	34
3.2	Performance of sentiment classifier	35
3.3	Performance of sentiment classifier: effect of image scrambling	37
3.4	Performance of emotional biconcept classifier	37
4.1	CNN architecture used for audio sentiment classification	45
4.2	Performance of audio sentiment classifier	45
6.1	Number of video clips	61
6.2	Precision of clips: overall	74
6.3	Precision of clips: various cuts	75
6.4	Precision of clips extracted from audio emotional arc: smaller confidence intervals are more precise	76
6.5	Precision of clips: genre	76
6.6	Precision of combined model	80
7.1	Emotional content features: coefficients and p-values for model	84
7.2	Emotional cluster features k=5: coefficients and p-values for model	87
7.3	Emotional cluster features k=8: coefficients and p-values for model	93
7.4	Emotional cluster features k=10: coefficients and p-values for model	94
B.1	Average valence of emotions	115
B.2	Percent of ratings that contain given emotion	116
B.3	Used to convey sentiment - full answer	117

B.4 Used to convey sentiment - individual items 118

Chapter 1

Introduction

“Facts don’t persuade, feelings do. And stories are the best way to get at those feelings.”

— Tom Asacker

Stories have long served as the throughline threading together hundreds of generations of humans. As social creatures, we have used stories to entertain, inform, and forge bonds. Theories behind the origin and purpose of stories are numerous, with interpretations including story as a form of “social glue”, escapist pleasure, practice for social life, and cognitive play [18, 17, 28].

However, not all stories are created equal. What can make one story draining, another story cathartic, and yet another story fall flat?

Can we understand such differences in response through the lens of *emotional arcs*? The existence of a core set of emotional arcs has long been discussed. Kurt Vonnegut once proposed the concept of “universal shapes” of stories, defined by the “Beginning–End” and “Ill Fortune–Great Fortune” axes. He argued nearly all stories could be categorized by the shape of its emotional arc, ranging from “Cinderella” (rise-fall-rise) to “bad and getting worse” (fall-fall) [2, 77]. Others have made similar claims, whether

it be 7 Jungian-derived arcs [15], 7 arcs defined by the nature of the antagonist [9], or 20 arcs defined by various drivers of plot [72].

There is, in fact, evidence that measuring stories along emotional dimensions can explain the degree to which a story engages its audience. In studying the virality of online content, the authors of [13] and [52] examined whether the valence, emotionality, and likelihood of eliciting specific emotions was predictive of New York Times articles making the Times' most e-mailed list. Ultimately, they found that emotional and positive media were more likely to be shared.

Perhaps not too long ago, the idea that a machine could watch a film, recognize emotions a human might, and use that experience to predict how engaging it is could sound highly improbable. But due to advances in neural networks and availability of data, we believe such a machine is increasingly within reach.

Motivated by the surge of video as a means of communication [35], the opportunities for machine modeling, and the lack of existing research in this area, we tackle these questions by viewing movies through emotional arcs.

1.1 Contributions

1. **Construction of audio and visual-based emotional arcs.** We use both existing and novel datasets to train deep convolutional neural network-based sentiment classifiers that can in turn be used to compute emotional arcs for movies for each of the two modalities. Critically, we also describe and use a crowdsourcing approach to evaluate the precision of key moments in these emotional arcs.
2. **A method for finding typical shapes of emotional arcs.** We motivate both a) the use of dynamic time warping as a metric to compare the shapes of emotional arcs, and b) k-medoids as a method to cluster the emotional arcs and

find typical shapes that reflect commonly occurring types of stories in movies. We find these arcs for high quality films created to tell stories, both of the feature-length and shorts variety.

3. **Analysis of how well emotional arcs predict audience engagement.**

We provide an example in which a movie’s emotional content features and emotional arc are statistically significant predictors of engagement, where we define engagement as the number of comments an online video receives.

1.2 Overview

Figure 1-1 illustrates the most salient pieces of the system, as well as how the pieces relate to each other. Here we also introduce some general concepts and terminology used to frame and describe this work.

Notably, we distinguish between modeling at two different scales – micro-level sentiment and macro-level emotional arcs. At the micro-level, we use visual and audio neural networks to predict sentiment on the order of seconds. The visual model takes a single image (frame) as input, while the audio model takes a 20-second sample as input. Modeling at the micro-level is in turn used to model at the macro-level, where the time scale is the length of the movie. An arc is composed using the sequence of predictions for seconds-level moments.

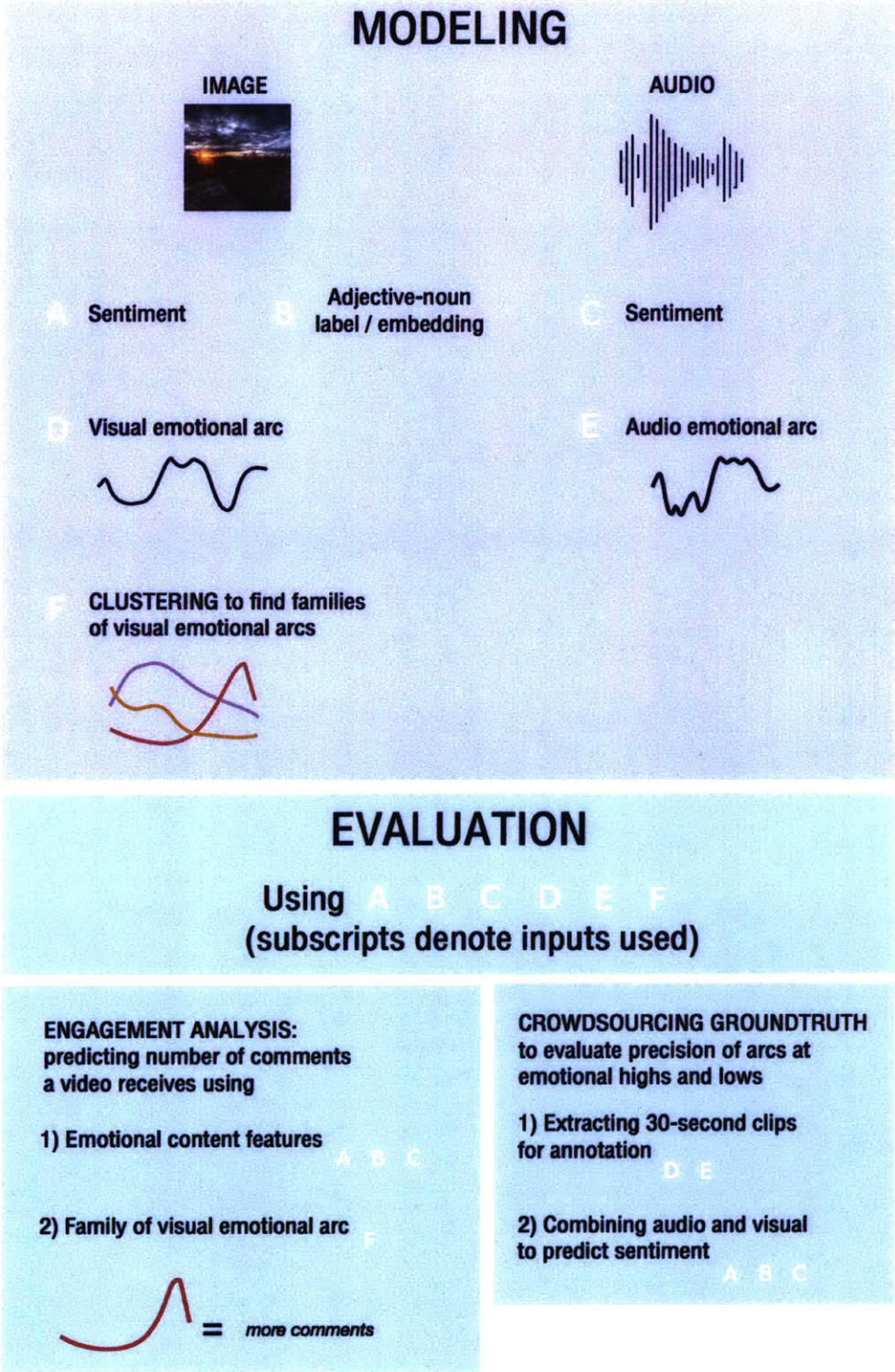


Figure 1-1: Overview

Reflecting this distinction, we evaluate each level separately and in parallel streams. First, we evaluate the system’s ability to accurately extract micro-level emotional highs and lows, which we sometimes refer to as *emotionally charged moments* or emotional *peaks and valleys*. This stream is shown on the right side of the evaluation section in Figure 1-1. Specifically, we measure precision as the amount of agreement in polarity between the micro-level models’ sentiment predictions and annotators’ ratings.

At the same time, the macro-level modeling is expanded upon and evaluated on the left side of Figure 1-1. First, we cluster the arcs into what we call *families* of arcs, each typified by a *typical* arc. These have colloquially been described in theories of narrative as “*core emotional arcs*” or “*universal shapes*” of stories. We may occasionally, in similar fashion, refer to arcs as shapes. These are in turn used by the subsequent engagement analysis.

Finally, we emphasize the main thrust of the work, which is to demonstrate that emotional arcs exist in movies and have predictive power for some measure of engagement, albeit in a small example. In Figure 1-1, the relevant stream of work is (A) -> (D) -> (F) -> Engagement Analysis 2, which ultimately shows that certain emotional arcs are statistically significant predictors of the number of comments a video receives.

1.3 Outline

The remaining chapters are arranged into parts, with Part I covering the modeling, Part II the evaluation, and Part III the conclusion.

Part I is divided as follows:

- Chapter 2 introduces the movie datasets for which we construct emotional arcs. We also describe the methodology used to create audio and visual emotional

arcs.

- Chapters 3 and 4 detail the datasets and architectures used to train audio and visual sentiment classification networks.
- Chapter 5 describes emotional arcs through typical arcs found through k-medoid clustering with dynamic time warping.

Part II is divided as follows:

- Chapter 6 evaluates the *micro-level* modeling by crowdsourcing ground truth data in order to a) measure the difficulty of the sentiment prediction task, b) evaluate the precision of emotionally charged moments in the arcs, and c) learn a model that combines the predictions from the audio and visual classifiers.
- Chapter 7 evaluates the *macro-level* modeling by using general emotional features and the categorization of emotional arcs in order to predict online engagement.

Part III is divided as follows:

- Chapter 8 discusses related work in emotional arcs, sentiment analysis, deep convolutional neural networks, and computational methods for understanding story.
- Chapters 9 and 10 conclude with limitations, suggestions for future research, and a discussion on the impact of the work.

Part I

Modeling - feeling from sights and sounds

Chapter 2

Constructing emotional arcs

In this chapter, we first introduce the two sets of movies for which we construct emotional arcs. We then provide an overview of how emotional arcs are created using our image and audio models, with chapters 3 and 4 providing details on the models themselves.

2.1 Video datasets

The system operates on two datasets collected for this work - the first consisting of Hollywood films, and the second consisting of high quality short films from the video-hosting website Vimeo. We selected films because they are specifically created to tell a story. This is in contrast to, for example, the YouTube videos often shared on social media. These stories are also *contained*, in contrast to serial television shows that would require modeling across episodes (though movie sequels do somewhat fall into this category). Finally, that there exist common film-making techniques to convey plot and elicit emotional responses from viewers suggests that we may be able to find typical shapes of emotional arcs.

Next, the motivations to include Vimeo shorts is threefold. First, it allows us to find

differences in storytelling that may exist between films and possibly newer, emergent formats unconstrained by studio demands. Second, modeling shorter form films could be the gateway to modeling, more generally, the short form videos that commonly spread through social networks. Finally, there is readily available engagement data for these shorts in the form of online comments left on Vimeo. This allows us to examine the predictive power of various emotional features, as shown in Chapter 7.

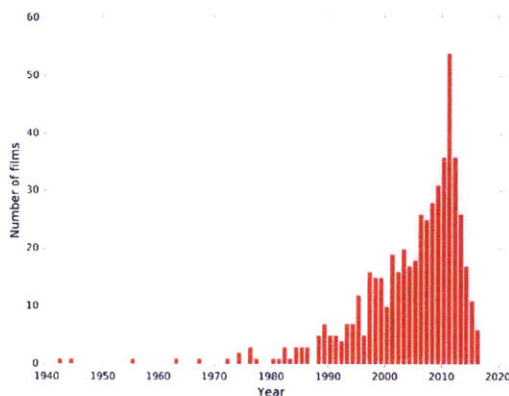


Figure 2-1: Films Corpora: distribution of release year

The first dataset is a collection of 509 Hollywood films, which we refer to from now on as the *Films Corpora*. The movies were released anywhere between 1944 and 2016, with the median release year being 2007. The distribution is shown in Figure 2-1. Notably, there is considerable overlap between this collection and the MovieQA [69] and M-VAD [73] datasets. This was a deliberate decision such that future work can build off of and incorporate the detailed captioning data included in these datasets.

The second corpora, which we call the *Shorts Corpora*, is a collection of 1,326 shorts collected from the Vimeo channel ‘Short of the Week’¹. These shorts are collected and curated by a team of filmmakers and writers. Each short is anywhere from 30 seconds to over 30 minutes long, with the median length being 8 minutes and 25 seconds. The distribution is shown in Figure 2-2.

¹<https://vimeo.com/channels/shortoftheweek>

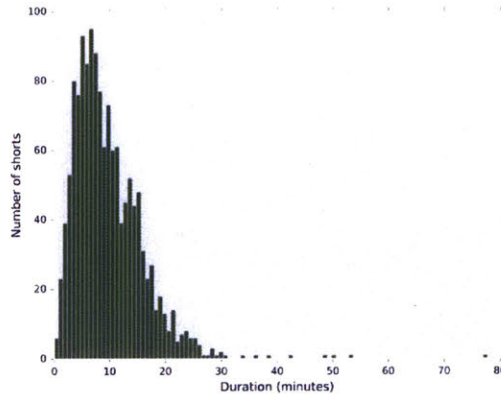


Figure 2-2: Shorts Corpora: distribution of duration

2.2 Methodology

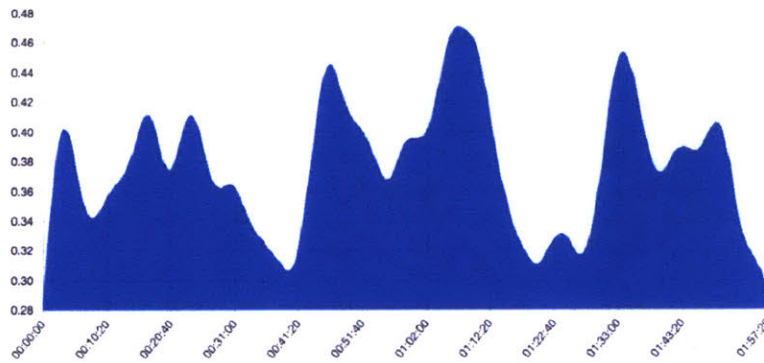


Figure 2-3: Visual emotional arc

Once the visual and audio models are trained for sentiment prediction (to be explained in Chapters 3 and 4), each can be applied separately across the length of the movie to create a visual emotional arc and audio emotional arc. To construct the visual emotional arc, we extract a frame from every second in the movie. Each frame is then resized and center cropped to size 256×256 and passed through the visual sentiment classifier. Similarly, to construct the audio emotional arc, we extract sliding 20-second windows from the video. Note that these windows can be overlapping, with a greater overlap leading to a smoother arc. In practice, a 50% overlap is used, with each window sharing 10 seconds with the previous window.

Both the visual and audio predictions can be quite noisy from frame to frame and

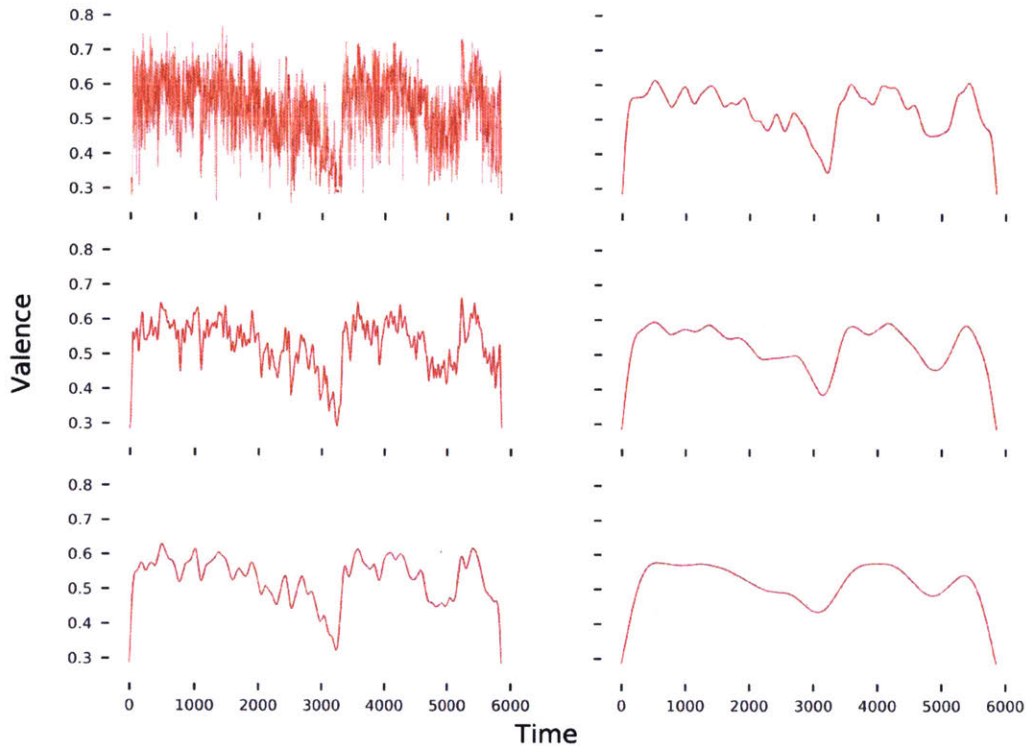


Figure 2-4: Emotional arcs: effect of smoothing values

Visual emotional arc for the movie *10 Things I Hate About You*. In the left column, the values of w are *none* (no smoothing), $0.01 * n$, and $0.03 * n$. In the right column, the values of w are $0.05 * n$, $0.1 * n$, and $0.2 * n$.

window to window, as temporal continuity is not an explicit part of the models. While the upper-left plot in Figure 2-4 shows that the macro-level shape is still visible from the raw predictions, it is helpful to smooth these signals to produce clearer emotional arcs. To do so, each 1D signal is convolved with a Hann window of size w , which is an approximately bell-shaped function. Figure 2-4 demonstrates the effect of various window sizes. In later downstream tasks, commonly used window sizes are 0.05, 0.1, and $0.2 \times n$, where n is the length of the video.

The audio and visual emotional arcs for the movie *Her* are shown in Figures 2-3 and 2-5. The audio emotional arc is also bounded by the confidence intervals to be described in chapter 4.

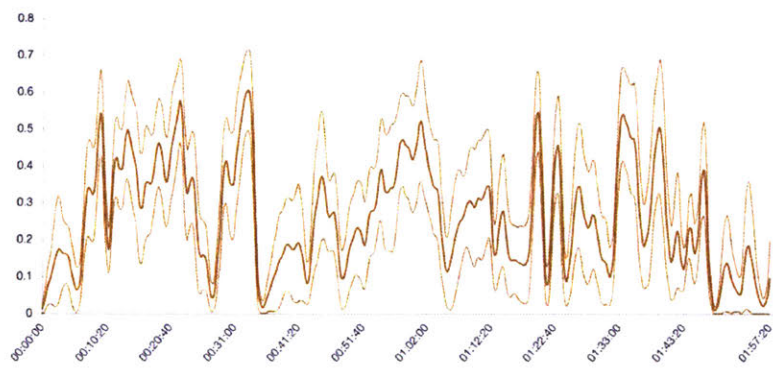


Figure 2-5: Audio emotional arc

Chapter 3

Image modeling

A video without visual is no video at all. Movies are watched, with light, color, framing, subject matter, and setting all playing a role in how the movie is perceived. The effect of various properties of the visual medium have been well studied, ranging from the positive psychological effects of nature scenes [74] to how color in movies is used to express story and emotion [1]. The primacy of color is so immediate and powerful that some filmmakers even employ colorscripts in pre-production, explicitly mapping color to target emotions [10]. With all this in mind, we built models that can take as input the core unit of film – the frames that compose it.

3.1 Sentibank dataset

Key to our work is the Sentibank visual sentiment dataset [16]. This dataset includes nearly half a million images, each of which is categorized according to an ontology of 1,533 adjective-noun pairs. There are a total of 213 adjectives and 356 nouns. These pairs, such as “charming house” and “ugly fish”, are termed *emotional biconcepts*. An example image for “dark clouds” is shown in Figure 3-1. Each biconcept is also mapped to an associated sentiment using an existing sentiment lexicon. We use these

values to train our sentiment classifier.



Figure 3-1: Example emotional biconcept

To construct the dataset, the authors used the 24 emotions found in Plutchik’s wheel of emotions as query terms to the photo-sharing website Flickr and video-hosting website YouTube. The most frequently associated tags for each emotion were then used to create candidate adjective-noun pairs. Pairs were then selected to create the final ontology according to a number of criteria, including but not limited to frequency, coverage, and specificity.

3.2 Model

We use a deep convolutional neural network to classify images, with the model architecture described in Table 3.1. This model is based on the AlexNet architecture, which was originally used to detect objects and scenes in the ImageNet challenge [42].

While an architecture that better reflects the state of the art would likely have higher accuracy, image-level classification accuracy was not the emphasis of our work. Instead, our focus was on building higher order arcs, for which this relatively simple model sufficed. However, we did make a few modifications to the original AlexNet design to reflect advancements since the time it was published. We briefly note each one and its purpose in the following:

- PReLU: parameterized version of the rectified linear unit (ReLU), both designed to combat the vanishing gradient problem, which can make training deeper neural networks difficult [32]. This was chosen over ReLU following experiments comparing different rectified activation functions in [78], as well as empirical results on tasks for this work.
- Batch normalization: the batch normalization layer speeds up training by normalizing the layer responses per mini-batch in order to reduce covariate shift. It also often improves generalization performance [36].
- ADAM: instead of using standard stochastic gradient descent to train our networks, we use ADAM, which computes adaptive learning rates using first and second order moments of the gradients [41].

Networks were implemented and trained using TensorFlow [8].

3.3 Sentiment prediction

To convert biconcepts to sentiment values, the authors used the SentiWordnet sentiment ontology. Each biconcept was mapped to a sentiment value as follows:

$$S(anp) = \begin{cases} S(adj) & \text{sgn}\{S(adj)\} \neq \text{sgn}\{S(noun)\} \\ S(adj) + S(noun) & \text{otherwise} \end{cases} \quad (3.1)$$

where $S(adj)$ and $S(noun)$ is returned by SentiWordnet, and sgn indicates the sign of the value. If the sentiment of the adjective and noun share the same polarity, then the sentiment of the biconcept is simply the sum of the two values. If they differ, however, the sentiment of the adjective is used, as the authors argue that the adjective is a stronger descriptive indicator of sentiment. Values in SentiWordnet are in the range -1.0 to 1.0. Thus, each adjective-noun pair is assigned a float value between -2.0 and 2.0. In order to train our classifier, we labeled all images with a sentiment greater

Group #	Layer
1	Conv: kernel= 3×3 , filters=96, stride= 4×4
	PReLU
	BatchNorm
	MaxPool: size= 3×3 , stride= 2×2
	Local response normalization: depth=4, bias=1.0, alpha=0.001 / 9.0, beta=0.75
2	Conv: kernel= 5×5 , filters=256, stride= 4×4
	PReLU
	BatchNorm
	MaxPool: size= 3×3 , stride= 2×2
	Local response normalization: depth=4, bias=1.0, alpha=0.001 / 9.0, beta=0.75
3	Conv: kernel= 3×3 , filters=384, stride= 1×1
	PReLU
	BatchNorm
4	Conv: kernel= 3×3 , filters=384, stride= 1×1
	PReLU
	BatchNorm
5	Conv: kernel= 3×3 , filters=256, stride= 1×1
	PReLU
	BatchNorm
	MaxPool: size= 3×3 , stride= 2×2
6	Fully connected: units=2048
	PReLU
	Dropout: prob=0.5
7	Fully connected: units=2048
	PReLU
	Dropout: prob=0.5
8	Fully connected: units=output dimension
9	Softmax
	Cross entropy loss

Table 3.1: Modified Alexnet architecture used for image classification

than 0.5 as ‘positive’, and those with a sentiment less than -0.5 as ‘negative’.

Training details for sentiment classification. The network was trained with a learning rate of 0.01, a batch size of 128, and a batch normalization decay of 0.9.

Results. The performance of the classifier is shown in Table 3.2. We note that the final accuracy is likely hurt in part by the noisiness of the dataset, which is a function of how it was created automatically. Figure 3-4 in the last section of this chapter

illustrates some of this noise. To address this problem, [79] introduces a framework for progressively fine-tuning networks by discarding noisy training examples.

Accuracy	Precision	Recall	F1
0.652	0.753	0.729	0.741

Table 3.2: Performance of sentiment classifier

3.3.1 Sentiment and color - image scrambling experiments

Color plays a large role in eliciting emotional responses [1, 10]. Imagine, for example, the simple difference between a dark room and the same one warmly lit. To better understand how much our network classifies based on objects versus color, we performed a simple experiment based on scrambling input test images.

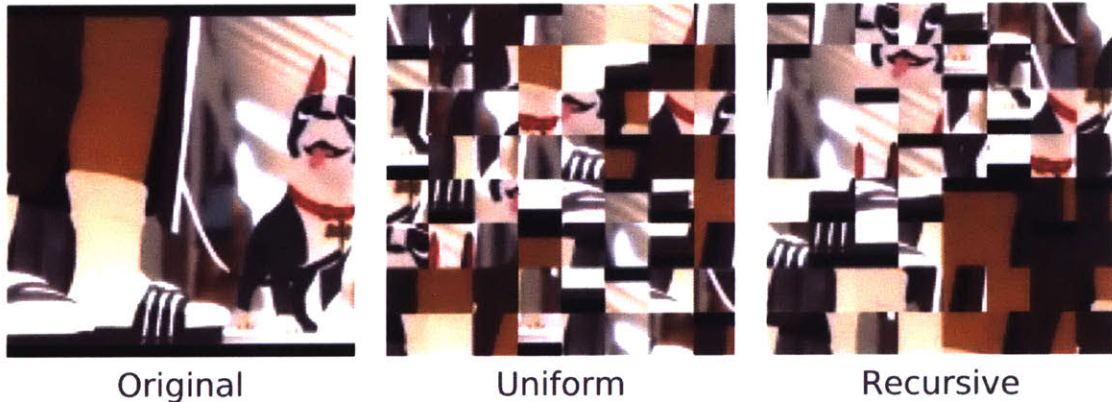


Figure 3-2: Image scrambling

We scramble an image of dimensions $n \times n$ by dividing it into blocks, and then permuting those blocks. We define two different methods for scrambling – uniform and recursive. Both scrambling functions take a parameter b , which defines the dimensions of one block. Here we assume n is divisible by b . In uniform scrambling, the image is divided into $(n/b)^2$ blocks, after which the blocks are randomly permuted. In recursive scrambling, we recursively subdivide an image until we reach blocks of size $b \times b$. After every recursive subdivision into four blocks, we randomly permute these four blocks. Examples of each method are shown in Figure 3-2, while Figure 3-3 illustrates the

effect of different block sizes. At a high level, recursive scrambling preserves more of the semantics of the image than uniform scrambling.

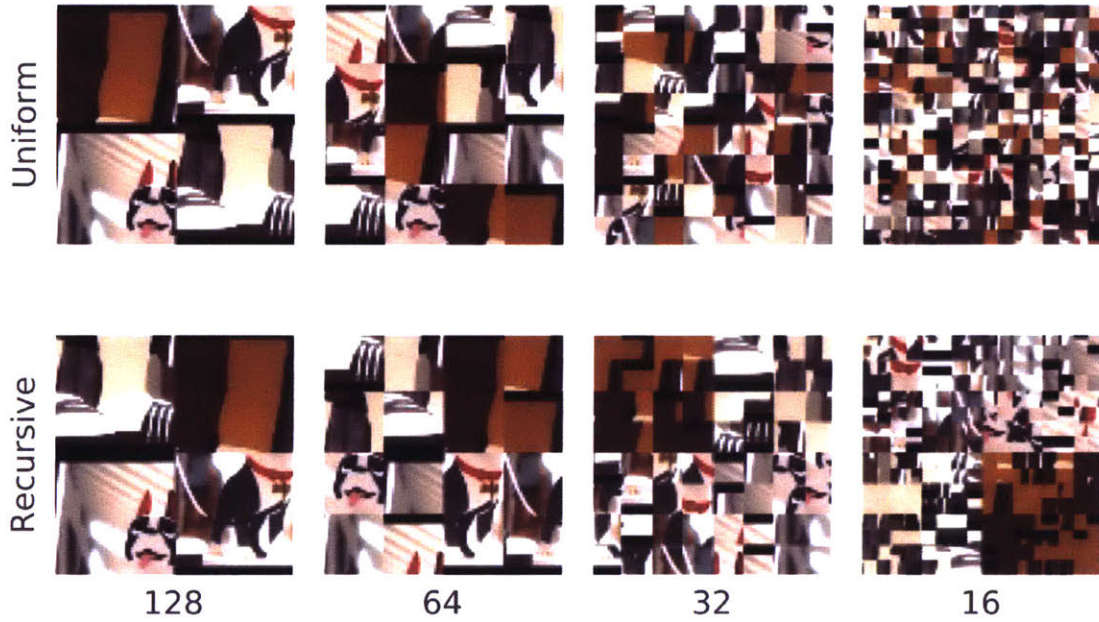


Figure 3-3: Image scrambling: varying block sizes

Results. Table 3.3 shows the effects on accuracy of scrambling a 224×224 image. As expected, when the block size is small (e.g. $b = 7$), performance is worse for uniform scrambling compared to recursive scrambling. On the whole, however, performance doesn't suffer *too* greatly once the block size is large enough. This suggests that the network is indeed largely classifying images based on color distribution features.

This is in part to be expected. First, splitting the data into only two classes (positive or negative) results in each class containing a extremely wide distribution of images. This may preclude the network from learning much beyond color and basic texture-based features. Second, we would expect a deeper neural network with a larger receptive field size to be more strongly affected by scrambling. To make a sentiment classifier that takes objects and scenes into greater account, we could fine-tune a network pre-trained for object or biconcept recognition.

Scramble mode	Block size	Accuracy	Precision	Recall	F1
Uniform	7	0.403	0.661	0.153	0.245
	14	0.616	0.674	0.704	0.687
	28	0.625	0.650	0.736	0.690
	56	0.627	0.671	0.650	0.660
	112	0.618	0.670	0.603	0.635
Recursive	7	0.557	0.669	0.478	0.558
	14	0.590	0.670	0.560	0.611
	28	0.619	0.671	0.644	0.657
	56	0.618	0.667	0.628	0.650
	112	0.617	0.671	0.600	0.637

Table 3.3: Performance of sentiment classifier: effect of image scrambling

3.4 Emotional biconcept prediction

While reducing the biconcepts to a single sentiment value is useful for creating emotional arcs, it also throws away a lot of information. Thus, we trained a second network that treats the biconcepts as labels. This network proves useful in creating ‘movie embeddings’ that capture in broad strokes a movie’s emotional content. Details are discussed in section 6.6.2.

Training details for biconcept classification. We used only biconcepts that had a minimum of 125 images, leaving 880 valid biconcepts. The network was trained with a learning rate of 0.01, a batch size of 128, batch normalization decay of 0.9, and a weight decay of 0.005.

Quantitative Results. The accuracy of the classifier is shown in a) of Table 3.4. Top- k accuracy is defined as the percent of images for which the true label was found in the top- k predicted labels. Again, labels are adj-noun pairs.

Top-1	Top-5	Top-10
7.4%	19.9%	28.4%

(a) Predicting adj-noun pair

Match	Top-1	Top-5	Top-10
Adj	12.0%	30.2%	41.5%
Noun	15.1%	31.7%	40.7%

(b) Matching adj / noun in predicted adj-noun pair

Table 3.4: Performance of emotional biconcept classifier

We also calculate top- k accuracy for adjectives and nouns. The predictions are still adj-noun pairs, but we calculate the percent of images for which the true adj/noun is found in the top- k predicted adj-noun pairs. Results are shown in b) of Table 3.4. The higher accuracy suggests that even though the predicted adj-noun pair may be incorrect, the label has learned both emotional and semantic relationships between images. This idea is reinforced in the subsequent qualitative analysis.

Qualitative Results. We also qualitatively examine the top predicted labels for a few images, as shown in Figure 3-4. The true label is beneath each image, the top 10 predicted labels are on the right of each image, and any correctly predicted labels are bolded.

The results show that even when the true label isn't found in the top 10 predicted labels, the predicted labels still largely returns semantically relevant biconcepts. This examination also serves to illustrate a number of things about the dataset. Specifically, we see:

- **Noisy labels.** For instance, one could argue “little beauty” hardly describes a spider, no matter how colorful it looks. The image for “young artist” also describes the work of the artist, not the artist themselves.
- **Abstract biconcepts.** It's not clear what set of images could accurately encapsulate a concept like “bad guy”.
- **Missing biconcepts.** Though “violent crime” is a biconcept, it's highly unlikely any image found on Flickr would actually depict the act of a violent crime. Likewise, we're unlikely to see images in this dataset related to concepts such as gore or death.



rough_sketch
 dead_fly
 broken_glass
 broken_mirror
 falling_angel
 cord_wor
 dead_bug
 sexy_shoes
 cord_jack
 falling_snow

young_artist



young_man
 traditional_dance
 nice_hat
 little_kid
 cute_kids
 charming_city
 traditional_dress
 silly_cat
 fancy_dress
 super_hero

bad_guy



famous_landmark
 christian_church
 ancient_church
 holy_cross
famous_church
 ancient_monument
 holy_spirit
 famous_lower
 ancient_city
 holy_mother

famous_church



cute_cat
 wet_cat
 cute_face
 adorable_kitty
 pretty_kitty
 sweet_kitty
 cute_kitty
 cute_animals
 curious_cat
 frenzied_cat

fluffy_ears



smiling_kids
 cute_kids
 happy_family
 proud_parents
happy_kids
 little_girl
 sweet_smile
 cute_smile
 young_girls
 cute_baby

happy_kids



wild_bird
 wet_leaves
 wet_grass
 little_bird
 traveling_birds
 dead_owls
 tiny_spider
 beautiful_bird
 wild_nature
 dark_eyes

little_beauty



empty_street
 quiet_street
 lonely_road
 abandoned
 empty_chair
 wet_road
 clean_car
 drunk_driver
 busy_street
 lost_control

violent_crime



hot_car
 classic_cars
 clean_car
 heavy_duty
 broken_car
 dirty_car
tiny_car
 broken_window
 empty_chair
 super_cars

tiny_car



traditional_dance
 nice_hat
 traditional_dress
 fancy_dress
 silly_hat
 evil_groom
 creepy_doll
 angry_halloween
 funny_hat
 holy_spirit

traditional_dance

Figure 3-4: Biconcept predictions

Chapter 4

Audio modeling

Imagine ‘watching’ a movie with your eyes closed – even limited in such a way, you would likely still be able to pinpoint moments of suspense, silliness, and sadness. Together with the visual events, audio is used in movies to to elicit emotional responses from viewers [58]. Sound and music can also play against the scene displayed, with the ear scene in Quentin Tarantino’s *Reservoir Dogs* and the Hip to Be Square scene in *American Psycho* as prime examples of how such contrast can heighten a scene. Audio is often secondary to the more obvious visual stimuli, and one can argue that an effective score floats just beneath the surface, arising at key moments to grab the viewer [6]. With the idea that just a few seconds is enough to set the mood, we created a model for sentiment classification that operates on 20-second snippets of audio.

4.1 Spotify dataset

We first started with the Million Song Dataset [14], which contains various metadata for approximately one million songs. Several of these features, such as tempo or major/minor key, could theoretically be used as a proxy for valence, energy, and other

relevant target features. We could also use the associated Last.FM dataset, which provides user-provided tags such as “happy” or “sad” for a subset of songs. Unfortunately, the subset of tagged songs is significantly smaller than the full dataset. We also note that the audio for the songs is not included in the Million Song Dataset.

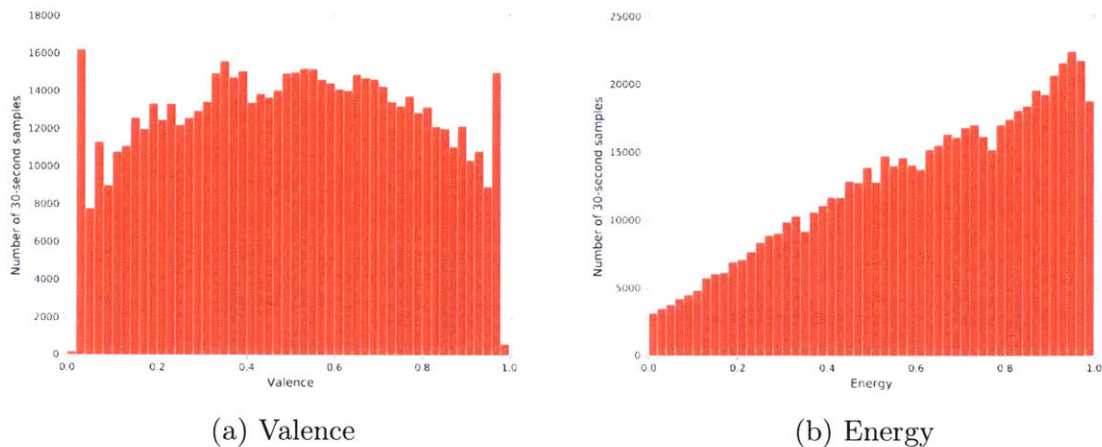


Figure 4-1: Spotify dataset: distribution of valence and energy features

To address our needs, we turned to the Spotify API. The API provides not only 30-second samples for songs, but also a number of audio features used by the company internally. These include *valence*, which measures the “musical positiveness conveyed by a track”, and *energy*, which measures the “perceptual measure of intensity and activity”. The distribution of valence and energy values are shown in Figure 4-1¹.

We used song titles from the Million Song Dataset as queries to the API and collected data for a total of 635,399 songs.

For the task of sentiment prediction, we use all samples that have a valence either greater than 0.75 or less than 0.25. This leaves a total of approximately 200,000

¹More from <https://developer.spotify.com/web-api/get-audio-features/>:

- “Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).”
- “Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.”

samples. While we don't train a network to predict energy in this work, we believe it could be a useful and complementary signal, especially in the context of movies. Consider a fighting scene in a superhero movie – the sounds may not be explicitly positive or negative, but it's quite possible that they have high energy. The energy could then affect the magnitude of excitation and arousal in the viewer. This also aligns with emotional frameworks that include pleasure (valence) and arousal (energy) [51]. Further suggesting that energy could be a non-redundant, complementary signal, Figure 4-2 shows the non-linear relationship between the two.

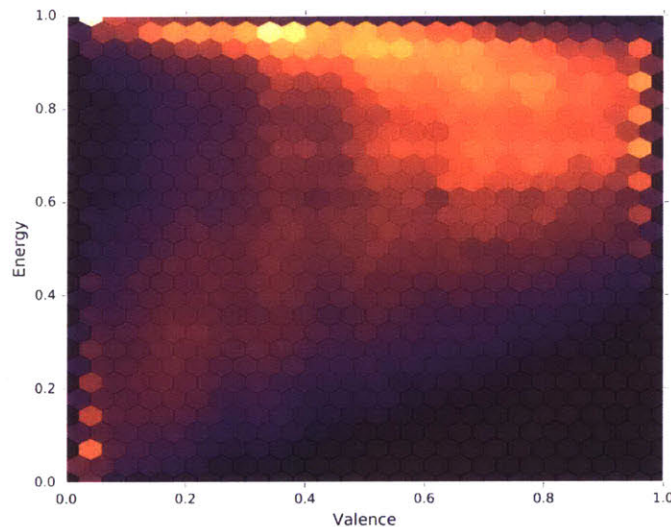


Figure 4-2: Spotify dataset: relationship between valence and energy. The brightness of each hexbin denotes the density of audio samples, with brighter values indicating greater density.

4.2 Audio representation

We represent each 30-second audio sample as a mel-spectrogram. Frequencies in the spectrogram are transformed to the mel scale according to equation 4.1. The result of perceptual listening experiments, the mel scale attempts to reflect the non-linear human sensitivity to different pitches. An example mel-spectrogram for a 20-second slice of audio is shown in Figure 4-3. We use 96 mel-bins for our models and the

Python package librosa to compute the mel-spectrograms [50].

$$m = 2595 \log_{10}(1 + f/700) \tag{4.1}$$

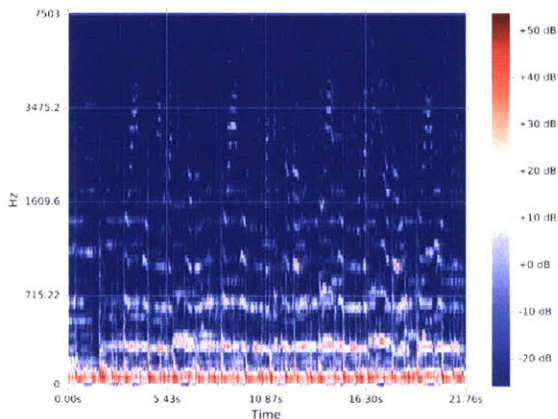


Figure 4-3: Example mel-spectrogram

We briefly note that mel-frequency cepstral coefficients (MFCC's) are another commonly used representation. Not too unlike mel-spectrograms, MFCC's are (roughly) created by taking windowed Fourier transforms, mapping onto the mel scale, and then taking the discrete cosine transform of the logs of the powers on the mel scaled frequencies [70]. [21] demonstrate that mel-spectrograms and MFCC's have comparable performance on a music tagging task.

4.3 Model

Once an audio sample is represented as a 2D mel-spectrogram, we can utilize the power of deep convolutional neural networks (CNN's). Following [21], we adopt the architecture shown described in Table 4.1. While it's possible to stack a recurrent neural network on top of a feature-extracting CNN in order to model the temporal aspect of audio, [22] shows this provides only marginal gains at the cost of increased computational complexity.

Group #	Layer
1	Conv: kernel=3 × 3, filters=32, stride=1 × 1
	ELU
	BatchNorm
	MaxPool: size=2 × 4, stride=2 × 4
2	Conv: kernel=3 × 3, filters=128, stride=1 × 1
	ELU
	BatchNorm
	MaxPool: size=2 × 4, stride=2 × 4
3	Conv: kernel=3 × 3, filters=128, stride=1 × 1
	ELU
	BatchNorm
	MaxPool: size=2 × 4, stride=2 × 4
4	Conv: kernel=3 × 3, filters=192, stride=1 × 1
	ELU
	BatchNorm
	MaxPool: size=2 × 4, stride=2 × 4
4	Conv: kernel=3 × 3, filters=256, stride=1 × 1
	ELU
	BatchNorm
	MaxPool: size=4 × 4, stride=4 × 4
6	Reshape
	Dropout, prob=0.5
	Fully connected: units=2
7	Sigmoid
	Cross entropy loss

Table 4.1: CNN architecture used for audio sentiment classification

4.4 Sentiment prediction

The performance of the sentiment classifier on the test set is shown in table 4.2. The accuracy is 89.6% and the F1-score is 0.900.

Accuracy	Recall	Precision	F1
0.896	0.871	0.931	0.900

Table 4.2: Performance of audio sentiment classifier

4.5 Uncertainty estimates

We face a problem when blindly applying a model trained on the Spotify dataset to audio found in movies. Namely, the audio found in movies will often contain sound not found in the song-based dataset. For instance, movies may have significant sections where the audio is background noise, conversation, or silence, to name a few possibilities.

This problem is an example of covariate shift, in which the distribution $P(x)$ for data modeled as being generated by $P(y|x)P(x)$ changes between training and test time [61]. One approach to handle ‘unfamiliar’ inputs is to explicitly predict if the audio comes from the training distribution. Another approach is to produce confidence intervals for every prediction. Both have the end goal of weighting predictions.

We sought a solution under the second approach. While the softmaxed activations from a neural network can be interpreted as probabilities, and hence a reflection of confidence, these probabilities can often be biased and require calibration [55]. Bayesian neural networks that can produce these uncertainty estimates by modeling parameter uncertainty is an active area of research.

We follow a method introduced in [26], which produces approximate uncertainty estimates for any neural network that contains dropout. Commonly used to prevent overfitting, a dropout layer with $prob = 0.5$ operates by ‘dropping out’ (setting to zero) each unit in the layer before it with probability 0.5 during training [67]. At test time, however, the dropout probability is typically 0. The proposed method bootstraps confidence intervals for an input point by setting the probability to 0.5 during test time, passing the point m times through the network, and using as the standard deviation of the predictions to define a confidence interval around the mean of the predictions. We can see an example of confidence intervals for the audio predictions on movies in Figure 2-5.

We briefly note that this method can also be applied to our image models. We should

expect that the standard deviations be smaller as there is less covariate shift. This is indeed the case – standard deviations reach as high as 0.2 for audio predictions but only about 0.05 for image predictions.

Chapter 5

Finding families of emotional arcs

With the machinery to create emotional arcs, we now seek to find the typical “shapes of stories” that Vonnegut and others have long theorized about.

Results shown in this section are based on the *visual* emotional arcs. The audio arcs can also be clustered into coherent families, but we believe it makes much more sense for a movie to be contoured by a visual-scape than an audio-scape. Again, audio tends to be more localized to specific moments in a movie.

5.1 Approach: clustering using k-medoids and dynamic time warping

A naive approach to clustering emotional arcs could be to simply use a popular clustering algorithm such as k-means [47] with a Euclidean metric to measure the distance between two arcs. K-means is an iterative parametric algorithm where the user specifies apriori the number of clusters k . The algorithm starts by randomly selecting k points as the *centroids* of k clusters. It then alternates between assigning each point (e.g. an emotional arc) to the nearest centroid and updating each centroid

as the mean of all the points assigned to it. However, this is a poor approach for our problem, for it fails in two ways:

1. Taking the *mean* of emotional arcs can fail to find centroids that accurately represent the shapes in that cluster. Figure 5-1 illustrates a pathological example of when this fails. The mean of first two signals, representing emotional arcs, has two clear peaks instead of one. While this is a contrived, pathological case, we argue that this is still a fundamental flaw in this approach.
2. On a related note, we shouldn't be operating in Euclidean space, as the Euclidean distance between two emotional arcs doesn't necessarily reflect the similarity of their shapes. Again, consider the first two plots in Figure 5-1. While these curves are clearly similar in shape, their Euclidean distance may be quite large.

With these limitations in mind, we turn to k-medoids [39] with dynamic time warping (DTW) [24] as the distance function. K-medoids operates very similarly to k-means – the main difference is that the centroid is replaced with the medoid. Namely, instead of using the *mean* of all the points in the cluster, we update the medoid as the point that is the *median* distance to all other points in the cluster.

Next, DTW is an effective distance function for measuring the difference between two time series that may operate at different time scales. It has traditionally been used in speech recognition tasks, where a given utterance has a similar characteristic waveform regardless who speaks it, but one person may speak faster or slower than another.

5.2 DTW formulation

Given two time series A and B both of length n , we first construct a $n \times n$ matrix M , where $M[i][j]$ contains the squared difference between A_i and B_j . To compute the

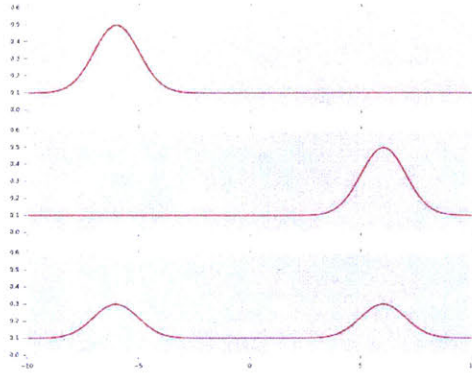


Figure 5-1: Pathological example of how k-means and Euclidean distance fails for clustering emotional arcs

DTW distance between A and B , we compute the shortest path through this matrix. This optimal path can be found in $O(n^2)$ using dynamic programming by:

$$c(i, j) = M[i][j] + \min\{c(i-1, j-1), c(i-1, j), c(i, j-1)\} \quad (5.1)$$

In Figure 5-1, for example, the DTW distance between the first and second time series is 0.

5.3 LB-Keogh for speed-up *and* better modeling

Several existing techniques to speed up DTW center around creating a ‘warping window’ that limits the available paths through the matrix M . Two commonly used windows are shown in 5-2. Using either of these windows, the popular Keogh lower bound approach creates upper and lower bound time series that envelop the original time series A as shown in Figure 5-3 [40]. Formally, these are defined as:

$$U_i = \max(A_{i-r} : A_{i+r}) \quad (5.2)$$

$$L_i = \min(A_{i-r} : A_{i+r}) \quad (5.3)$$

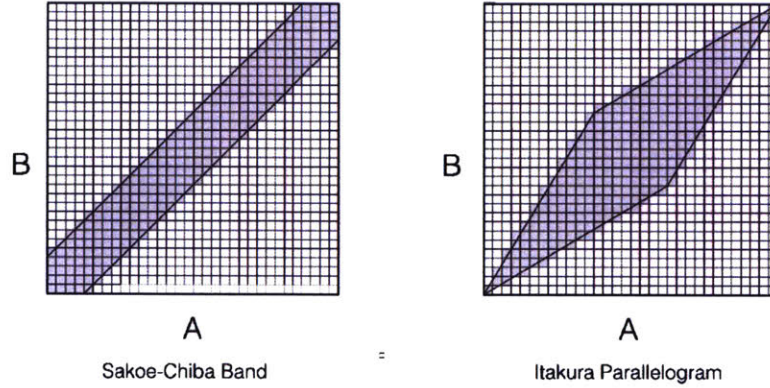


Figure 5-2: Warping windows for Keogh lower bound

This figure was adapted from [40].

where r is a parameter *reach* that controls the size of the window. Intuitively, this controls how much a time series is allowed to warp. The lower-bounded LB_{Keogh} distance between A and B is then given as:

$$LB_{Keogh}(A, B) = \sqrt{\sum_{i=1}^n \begin{cases} (B_i - U_i)^2 & \text{if } B_i > U_i \\ (B_i - L_i)^2 & \text{if } B_i < L_i \\ 0 & \text{otherwise} \end{cases}} \quad (5.4)$$

If B lies inside the envelope of A , then the distance is 0. If B lies outside the envelope, however, then its distance is essentially the distance to the nearest orthogonal edge of the envelope.

Importantly, this approach has ramifications beyond increased speed up. Consider again the first two emotional arcs shown in Figure 5-1. While they do have very similar shapes, both being characterized by one large peak, one could argue that they might impact a viewer very differently. The first hits the viewer with an emotionally charged moment as soon as the movie starts. The rest of the movie, however, is flat. The second, however, delivers the emotional blow at the end of the film – a moment timed in such a way could have a greater impact. Consequently, we would like our

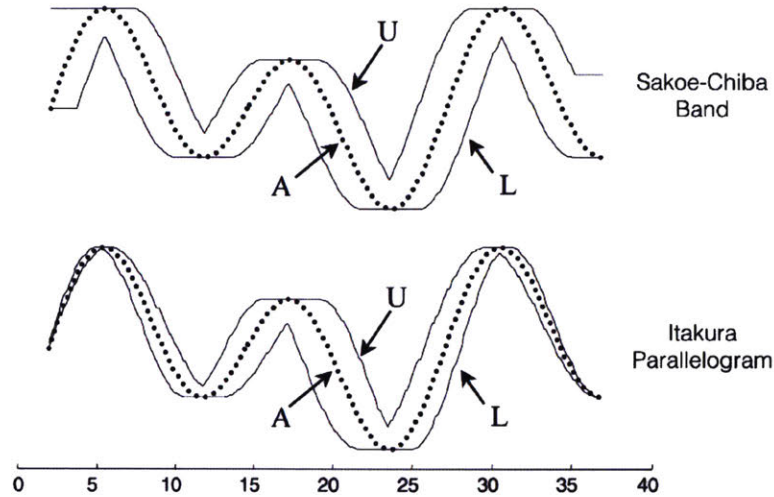


Figure 5-3: Envelopes for Keogh Lower Bound

This figure was adapted from [40].

distance function to only allow warping to a certain extent. The window does exactly that.

5.4 Practical notes

There are several practical considerations when applying DTW to the emotional arcs generated for our videos. First, the above formulation considers the distance of two time series both of length n . In practice, our movies are of different lengths. To account for this, we set a maximum length m for each corpora with $m = 10000$ for the Films Corpora and $m = 1800$ for the Shorts Corpora. Next, each video is stretched to length m , with missing intermediate values being interpolated. While it's capable to formulate DTW for time series of different lengths, experiments on a variety of tasks in [63] show there is little empirical evidence that either approach has a statistically significant effect on accuracy.

A few movies in each corpora exceed m and are excluded from the clustering analysis. Rather than a limitation, these videos can be considered outliers that should indeed be withheld. For instance, considering a 80 minute video in the Shorts Corpora as

a ‘short’ makes little sense. Including it would also force every movie, such as one lasting two or three minutes, to be stretched to 80 minutes.

Second, since we are interested in the overall shape of the emotional arc, we z-normalize each emotional arc such that all values lie in the range [0,1]. Specifically, a given emotional arc A , whose original values representing the valence predictions may lie in the range [0.4, 0.8], is normalized by the following:

$$A_{range} = \max(A) - \min(A) \tag{5.5}$$

$$A_i = (A_i - \min(A)) / A_{range} \tag{5.6}$$

Finally, we consider how to select r . $0.1 * n$ was commonly used in the speech recognition community, though recent experiments indicate that value was historical rather than practical. After testing a few different values, we settled on $0.025 * n$. This is close to $0.03 * n$, the value found to be optimal for a number of different tasks [63].

5.5 Cluster results

Though there is no ‘right’ number of clusters, there are a number of techniques for finding *an* optimal k . We use the elbow method [29], which plots the within cluster distance defined in Equation 5.7 against the number of clusters k . We then look for ‘elbows’ in the curve where the curve ‘flattens’ out. In Figure 5-4, for example, we can see possible ‘elbows’ at $k=5$ and $k=9$. We also briefly note that an experiment with k-means resulted in elbow plots with no clear decrease in the within cluster distance, let alone a discernible elbow. This reinforces our decision to use k-medoids and DTW.

$$WCD = \sum_{i=1}^K \sum_{x \in m_i} dist(x, m_i)^2 \quad (5.7)$$

where m_i is the i -th medoid, $x \in m_i$ signifies x belongs to the cluster with m_i medoid, and $dist(x, m_i)$ is the DTW distance between x and m_i .

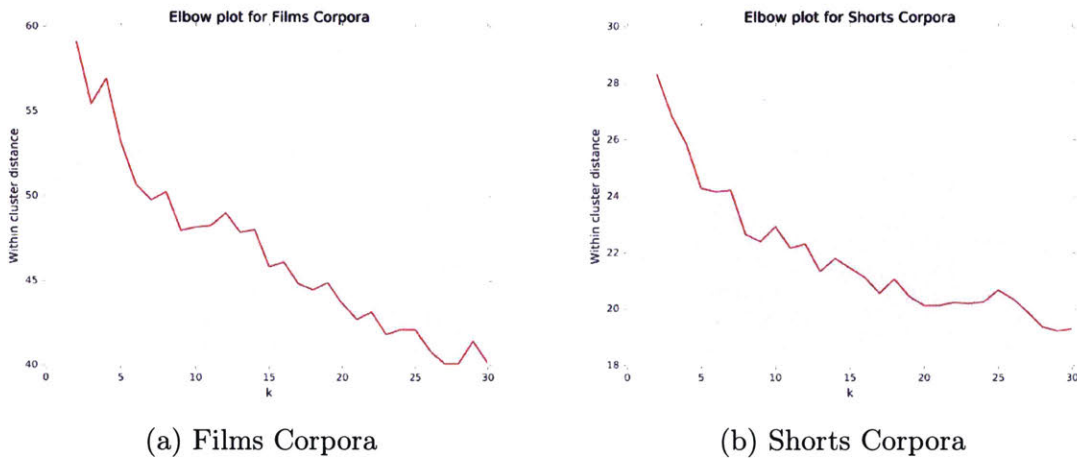


Figure 5-4: Elbow plots for k-medoid clustering

We show one example clustering for $k=5$ and $w=0.1$ in Figure 5-5 and 5-6. These arcs represent five typical emotional arcs. We note that the steep inclines and declines at the start and end of arcs are simply artifacts of a movie's opening and closing scenes and credits. When comparing Film arcs to Short arcs, we not surprisingly see that Short arcs are less complex. However, we also note more extreme arcs in the Shorts Corpora, such as the yellow arc that ends on a steady, depressing decline.

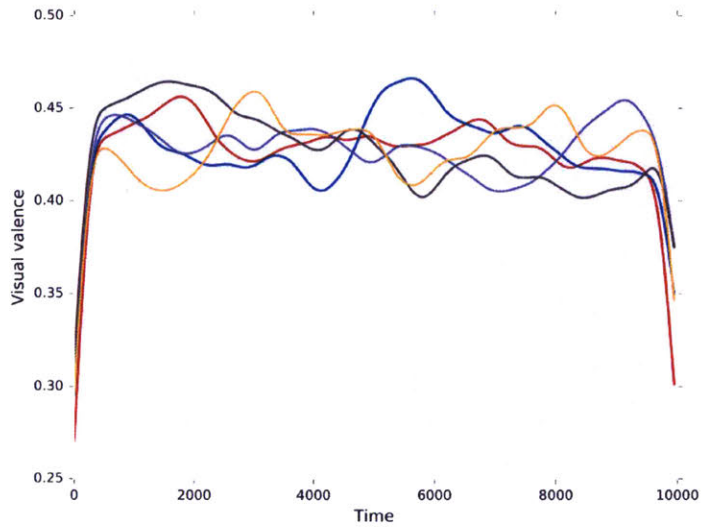


Figure 5-5: K-medoids on Films Corpora for $k=5$

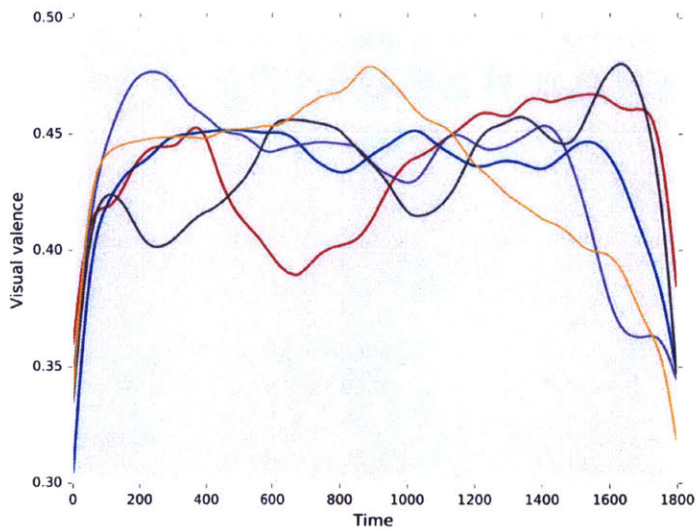


Figure 5-6: K-medoids on Shorts Corpora for $k=5$

Part II

Evaluation - moments and arcs

Chapter 6

Crowdsourcing ground truth

In this chapter, we evaluate the micro-level accuracy of our system and its ability to extract emotionally charged moments from a movie. In order to do so, we had to first collect ground truth data on movies. Having ground truth is critical because it allows us to:

1. **Understand the difficulty of the task.** Sentiment is subjective, and the accuracy of the system should be baselined against how much human responses may vary and contradict each other.
2. **Evaluate the accuracy of our audio and image sentiment models.** This is measured through precision, where we calculate the amount of agreement in polarity between our sentiment predictions and annotators' ratings.
3. **Combine the audio and visual predictions.** We can learn how to weight the two modalities. For example, if the predicted audio valence is high but the predicted visual valence is low, what is the likely actual valence?

The constructed emotional arcs were initially sanity checked and qualitatively assessed by having people a) visually inspect emotional arcs of movies they were familiar with, and b) compare moments in the audio and visual emotional arc that were high or low

against the actual video and audio at that moment in the movie. While this suggested that the system had some level of accuracy, it was far from quantitative.

One particularly thorough approach could be to have n subjects watch m videos and record their physiological responses throughout the viewing. While we briefly considered this with $n = 1$, due to time constraints, we looked to another approach.

Instead, we turned to crowdsourcing to annotate 30 second video clips extracted from various movies in the Films Corpora. Annotating an entire film, let alone multiple films, would be too cost and time intensive. Therefore, to keep the cost down and maximize the value of each annotation, we extracted video clips at high and low moments in the emotional arcs. We hoped that these moments, as opposed to scenes from more neutral moments, would lead to more interesting and informative annotations.

Workers were asked to watch a clip and answer four questions regarding its sentiment. We also highlight that each video clip was annotated by three workers in order to assess the difficulty of the task.

6.1 Video clip extraction

Overall, anywhere from one to seven clips were extracted from 168 films. In this section, we describe the methodology and design decisions that we used to select clips.

We refer to local maxima in an emotional arc as *peaks* and local minima as *valleys*. Combined with the two modalities, this gives us four types of clips – *audio-peaks* (a peak from the audio emotional arc), *audio-valleys*, *visual-peaks*, and *visual-valleys*. We extracted a roughly equal number of each, with a target of approximately 1000 clips total. Table 6.1 shows the exact numbers.

A common framework that underlies several methods for peak detection in signal

	Number of clips
Audio-peak	220
Audio-valley	253
Visual-peak	230
Visual-valley	259

Table 6.1: Number of video clips

processing literature is to first a) smooth the signal, and then b) find peaks using a filter, auto-correlation technique, or simple outlier detection method.

In our case, the emotional arcs were smoothed using the Hann window of size w . We use a window size of $w = 150$ for the audio emotional arcs and $w = 600$ for the visual emotional arcs. The w for audio is smaller to reflect the idea that audio in movie acts as a sharper tool for eliciting emotional responses and hence operates at finer time scales.

In our case, we find peaks by an auto-correlative method that compares a signal x to x^{-1} , where x^{-1} is x shifted by one. Peaks are then detected using the difference between x and x^{-1} , which essentially compares a point x_i against its neighbor. In addition, there are several tunable parameters, including a threshold for ‘peakiness’ of a point relative to its neighbors, ‘peakiness’ relative to the entire signal, distance between peaks, and flatness of peak (accounting for more neighbors). Figure 6-1 shows an example of peak detection.

To slice out a 30-second clip from a peak at point x_i , we take the segment $x_{i-15:i+15}$. In other words, we take the 15 seconds on either side of the peak.

Peak detection is performed on the audio and visual emotional arcs separately. In the infrequent but interesting case that the 30-second segment around a peak/valley in the audio arc overlapped with the 30-second segment around a peak/valley in the visual arc, we defined the joined extremum point as the midpoint of the overlap. A 30-second clip is then extracted centered around that midpoint, as shown in Figure 6-2. For simplicity’s sake, the program is designed to discard any cases that had 2

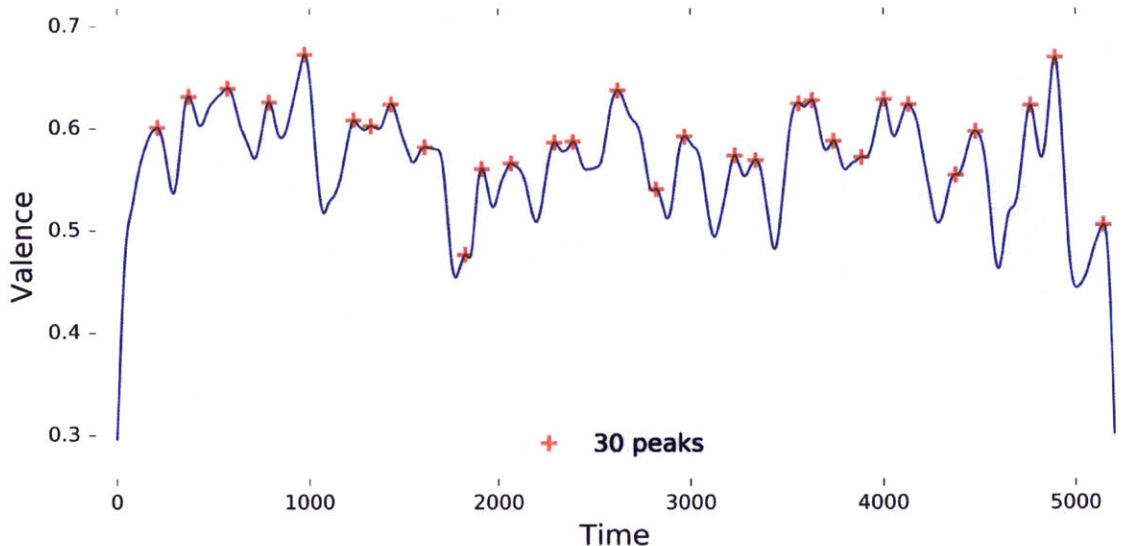


Figure 6-1: Peak detection example on the visual emotional arc of *Fantastic Mr. Fox* or more overlaps, as there is ambiguity around what time best to extract from. In practice, no such cases occurred.

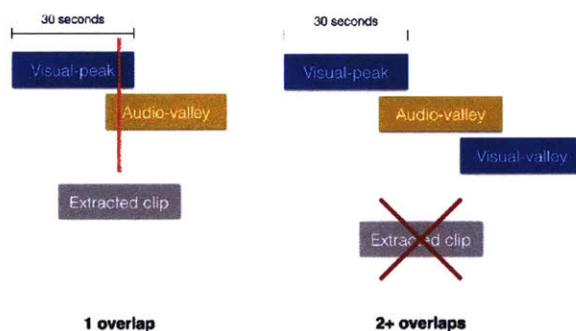


Figure 6-2: Clip extraction edge case: overlapping peaks/valleys

Finally, when extracting clips from audio-peaks or audio-valleys, there is the added consideration of how to sample with respect to the confidence intervals described in chapter 4. We hoped to show that predictions with smaller confidence intervals were more accurate. In other words, peaks/valleys with small confidence intervals would be more emotionally charged than peaks/valleys with larger confidence intervals. In order to do so, we needed a healthy sample size at varying confidence interval bins. Because there were relatively few audio-peaks with small confidence intervals, we performed biased sampling by taking as many clips from small confidence interval

audio-peaks as possible. Even so, the distribution was non-uniform, with, for example, 11 clips with standard deviations (which define the confidence interval) in the range [0.02, 0.04), and 68 clips with standard deviations in the range [0.08, 0.1). The exact numbers are shown in Table 6.4.

6.2 Crowdsourcing experiment

We chose to use the CrowdFlower platform for its simplicity and ease of use. While Mechanical Turk, another popular crowdsourcing platform, has greater flexibility in customizing an experiment, our task was relatively straightforward. Moreover, studies comparing the two indicate that the quality of annotations can be comparable [25].

6.2.1 Instructions and questions

6.2.1.1 Task

In our annotation task, users are first provided with a set of instructions and guidelines. Next, they are given five test questions that serve both to assess the quality of the worker and give the worker some practice examples. Once completed, they can move on to annotating clips.

6.2.1.2 Instructions

Overview

After watching a 30-second video clip, you will be asked to answer questions about the sentiment and emotions surrounding the video.

Rules & Tips

- Watch the entire video with sound on.
- Please try to make your decisions based on the overall video clip, as opposed to just the beginning or end.
- As you watch, please take note of the cues (auditory, dialogue, visual) that help you make your decisions.
- Decisions should be based only on the current video clip – each video clip is meant to be independent.

Guidelines:

A *positive* video clip may depict or be described by:

- acceptance
- awe
- happiness
- hope
- humor
- inspiration
- joy
- love
- peace
- pride
- relief
- serenity
- trust

A *negative* video clip may depict or be described by:

- anger
- anxiety
- apprehension
- disgust
- fear
- sadness

6.2.1.3 Questions

1. How positive or negative is this video clip? (1 being most negative, 7 being most positive)
2. How confident are you in your previous answer? (1 being least confident, 10 being most confident)
3. Which emotion(s) does this video clip contain or convey? (check all that apply or none of the above)
 - Options: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, none of the above
4. Which of the following contributed to your decisions? (check all that apply)
 - Options: audio, dialogue, visual (actions, scene, setting)

6.2.1.4 Design considerations

Q1. Aiming to cover common positive and negative emotions, the words used to characterize ‘positive’ and ‘negative’ in the guidelines were manually selected from a combination of [3] and [60]. The phrases ‘depict’ and ‘be described by’ were chosen to cover both the emotional content and how the video was conveyed emotionally.

Q2. In addition to using the inter-annotator agreement on each clip, we used this question to probe into the difficulty of the task.

Q3. The 8 emotions were selected from Robert Plutchik’s theory of emotion [60].

Q4. This question was included in part to assess the importance of the audio signal versus the visual signal. Moreover, this question allows us to get a sense of the gains that could be achieved by explicitly modeling dialogue.

6.2.2 Experimental setup and stats

6.2.2.1 Stats

Workers and judgments. A total of 100 unique workers participated in the job. The number of annotations per worker is shown in Figure 6-3. A total of 2,930 annotations were collected.

Demographics. Only workers in the United States were allowed to participate in the job. Each worker's state and city are known, but this information was not used for any analysis.

Time. To assess the clarity of the guidelines, efficacy of golden ticket clips, and appropriateness of the payment per annotation, an initial task with 47 clips was launched. 143 annotations were collected in 2 hours. The remaining 2,787 judgments were collected 1 day later in 27 hours.

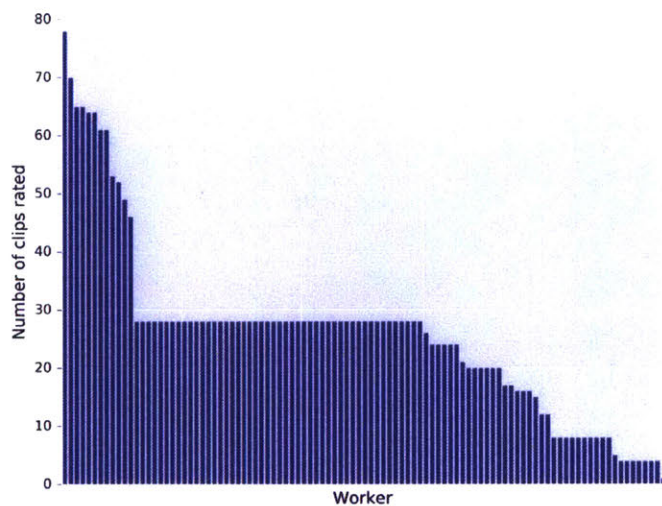


Figure 6-3: Number of clips annotated per worker

Annotators in CrowdFlower are also asked to evaluate the annotation task. The average scores are shown in the following list. We lack data for ratings on comparable tasks, but we provide these stats for posterity and as possible reference for others

designing CrowdFlower experiments.

- Overall: 4.7/5
- Instructions Clear: 4.6/5
- Test Questions Fair: 4.3/5
- Ease of job: 4.5/5
- Pay: 4.4/5

6.2.2.2 Quality control

The labels provided by workers on crowdsourcing platforms can vary greatly in quality. In addition to placing careful consideration upon the task instructions, one large lever in attracting workers and increasing quality is paying enough per annotation. In this section, we document the other methods used to assess worker quality.

Quality-based worker tiers. CrowdFlower allows one to filter workers from three different quality-based tiers [7]. A worker's tier level is based on the channel the worker comes from, his or her experience, and his or her past performance on other jobs. Allowing workers from tier 1 casts a wide net - your job will likely finish faster, but the quality may suffer. We allowed only workers from tier 3, the highest available tier.

Gold standard clips. Even after filtering by tier, it is prudent to actively assess the performance of annotators while the job is live. CrowdFlower allows further quality control through the use of 'test questions'.

In our case, 'test questions' are clips for which we were certain were highly positive or highly negative. These clips are here to referred to as *gold standard clips*. When a worker first joins the job, they are presented with five such clips. If they failed to correctly rate at least 3/5 clips, they weren't allowed to continue to the actual job. Here, 'correctly rate' means that the polarity of their valence rating must match the

clip's given ground truth. For instance, a 'positive' golden standard clip must be rated either 5, 6, or 7 (out of 7) to be considered correct.

We acknowledge that this is a rather loose assessment. However, asking for greater agreement (e.g. only 6 or 7) on this question is infeasible given the subjectiveness of the task. In the same vein, using question 3 (which emotions does this clip contain) is even more subjective and impractical.

To create an initial set of golden standard clips, we manually watched a number of clips and attempted to select ones that were highly positive or highly negative. Annotators are allowed to provide feedback or dispute test questions they disagree with. A handful of golden standard clips that received ambiguous feedback, such as both a positive rating and a negative rating from different annotators, were removed from the pool. This process further highlighted the subjectiveness of the task. Even clips that one might think is unequivocally positive could be considered negative by someone else.

After the job started running, clips that received strongly negative or strongly positive mean ratings were added as gold standard clips. A total of 24 gold standard clips were used.

Taking the job seriously. Finally, we attempted to filter out workers that weren't exercising due diligence. Annotators are shown a page with five clips at a time. As each clip is 30 seconds long, a page should require a minimum of 150 seconds to complete. Accounting for a few seconds per clip to assess the video and answer the questions, we set the expected minimum to be 175 seconds. Any annotator who completed a page in less time was automatically removed from the job and their annotations discarded.

6.3 Definitions and terminology

Here we define some terms that will be used when describing the results. These are in addition to the previously mentioned:

- **Peak.** A clip extracted from a high moment in an emotional arc.
- **Valley.** A clip extracted from a low moment in an emotional arc.
- **Audio-peak.** A clip extracted from a high moment in an audio emotional arc.
- **Audio-valley.** A clip extracted from a low moments in an audio emotional arc.
- **Visual-peak.** A clip extracted from a high moments in a visual emotional arc.
- **Visual-valley.** A clip extracted from a low moments in n visual emotional arc.

The following terms will also be used:

- **Valence rating.** An annotator's answer to Q1, rating the valence of the clip on a scale of 1 to 7.
- **Mean valence rating.** The mean of the three answers to Q1.
- **Positive/negative rating.** A positive rating is an answer to Q1 that is either 5, 6, or 7. Similarly, an answer of 1, 2, or 3 counts as a negative rating.
- **Positive/negative clips.** A positive clip is one in which the mean of the three valence ratings is greater than four. Similarly, a negative has mean valence ratings less than four.
- **Neutral clips.** Loosely counted as clips with a mean valence rating around 4.
- **Confidence rating.** An annotator's answer to Q2, regarding how confident they are in their answer to Q1.
- **Confidence of a clip.** The mean of the three answers to Q2.

6.4 Difficulty of task: inter-annotator agreement and confidence of ratings

To assess the difficulty of this labeling a video clip with sentiment, we look at three results.

6.4.1 Ambiguous clips

Recalling the definitions of positive and negative ratings in section 6.3, we define *ambiguous clips* as clips that received both a positive rating and negative rating. Under this definition, **16.99%** of all clips were ambiguous.

6.4.2 Variance of valence ratings

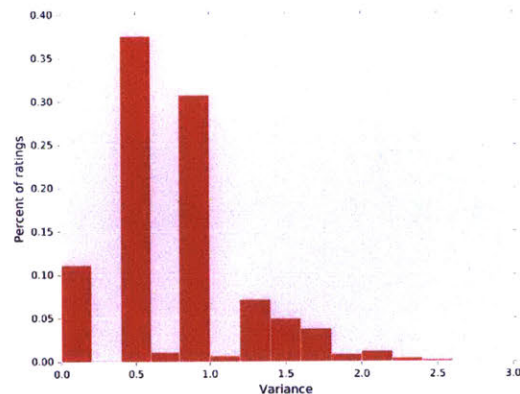
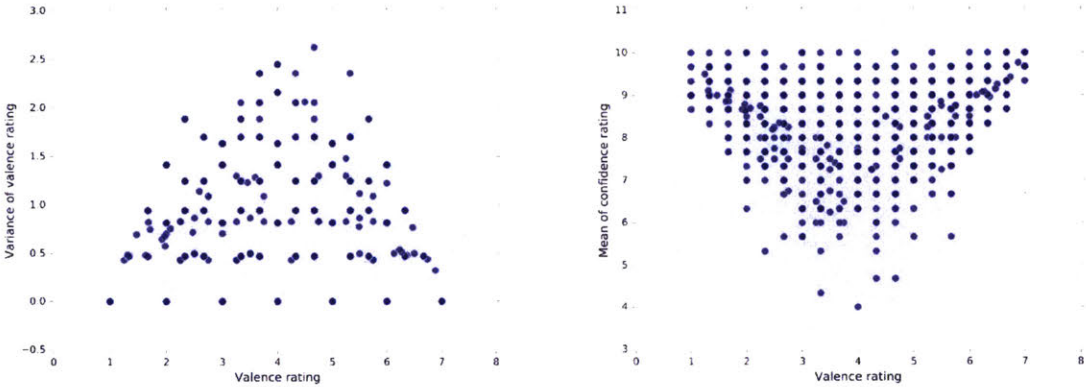


Figure 6-4: Variance of valence ratings

We also looked at the variance of the valence ratings of all clips. The variances are plotted in Figure 6-4, with the average variance across all clips being 0.8932. We note that the gaps in the distribution are simply a function of both the sparsity of annotators per clip and the small range of discrete answers possible. With more annotations, the roughly log-normal shape might be clearer. Also, there is a noticeable long tail of clips with relatively higher variance.

Next, for every clip, we compare the mean valence rating to the variance of the valence ratings. We would expect highly positive or highly negative clips to have relatively lower variance, as these clips may be more universally positive or negative. Likewise, we would expect neutral clips to have greater variance. This is exactly what occurs, as shown in plot (a) of Figure 6-5.



(a) Mean valence rating vs. variance of valence ratings (b) Mean valence rating vs. mean confidence ratings

Figure 6-5: Q2: Mean valence ratings, variance of valence ratings, and mean confidence ratings

6.4.3 Confidence ratings

Finally, we looked at the responses to Q2. Plot (a) of Figure 6-6 shows that workers are highly confident, with approximately 75% of all answers being 8 or 9. Because there are 10 to 20 users who made twice as many judgments as other workers (see Figure 6-3), one might wonder whether the skew towards high confidence is a function of this highly active subset of workers being unusually confident. To investigate this, we plot the distribution of the mean confidence *per worker* (i.e. the average of all of that worker’s responses to Q2). Plot (b) of Figure 6-6 shows that confidence ratings still skew towards the upper end.

In addition, for every clip, we compare the mean valence rating against the mean confidence rating. We would expect highly positive or highly negative clips to have

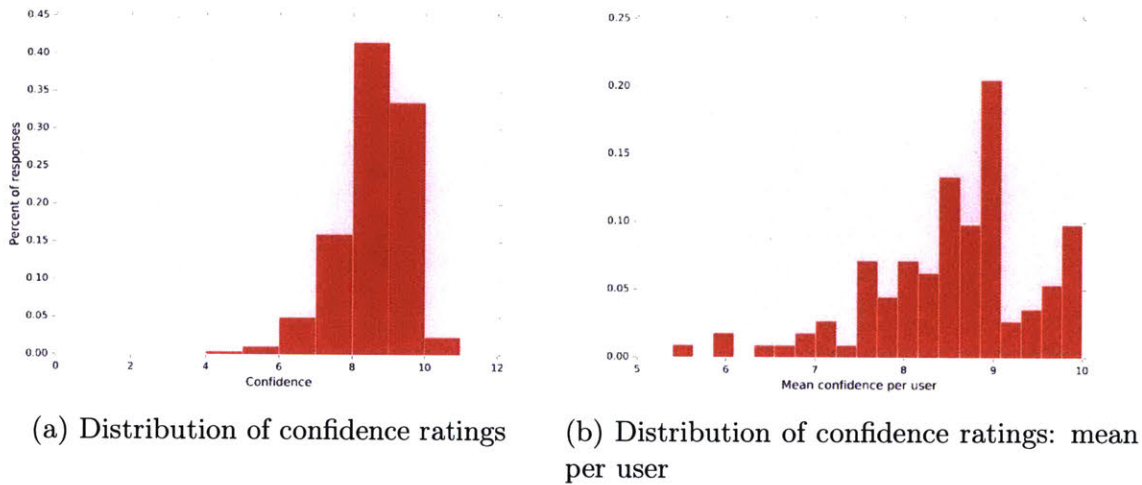


Figure 6-6: Q2: Confidence ratings

greater confidence. Likewise, we would expect neutral clips to have lower mean confidence ratings. This is confirmed in plot (b) of Figure 6-5.

Finally, we note an interesting relationship between the variance of valence ratings and the mean confidence rating. One might expect confidence to decrease as the variance of valence ratings increase. In a counter-intuitive result, Figure 6-7 shows that as the variance of the valence ratings increase, so too does the confidence. One hypothesis posits that clips with higher variance receive ratings at the far ends of the spectrum. One annotator might rate a clip a 6, while a second annotator might rate it a 3. In this case, the first annotator is giving a highly positive rating. If they have such a strong emotional reaction, they might be likelier to be extremely confident, thus raising the mean confidence rating.

6.5 Accuracy of system

6.5.1 Defining precision

With ground truth data in hand, we would like to assess the precision of our system.

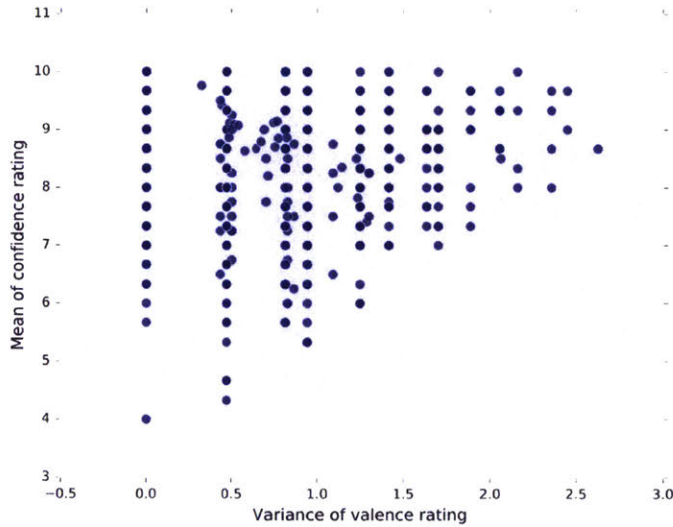


Figure 6-7: Variance of valence vs. Confidence

Recalling that every clip was extracted from a peak/valley in either the audio or visual emotional arc, we define the overall precision as:

$$\frac{(\text{peak \& positive}) + (\text{valley \& negative})}{\text{all clips}} \tag{6.1}$$

In other words, a clip was ‘accurately’ extracted if a) it was extracted from a peak in either emotional arc, and it was labeled as a positive clip, or b) it was extracted from a valley in either emotional arc, and it was labeled as a negative clip.

6.5.2 Overall precision

In table 6.2, we list the precision on both the full dataset and the full dataset with ambiguous clips removed. We note that random chance would be $3/7 = 0.429$.

We argue that ambiguous clips should not be included in the evaluation, as it is unclear what their valence might be without more annotations. As such, all further numbers are calculated on the set with ambiguous clips removed. However, we note that later analyses exhibit a similar difference of four to five percentage points between

the two sets.

Set	Precision
All	0.642
No ambiguous clips	0.681

Table 6.2: Precision of clips: overall

6.5.3 Precision by various cuts

We can also calculate the precision on various subsets of the data. We look at peaks versus valleys, audio versus visual, and the cross between the two.

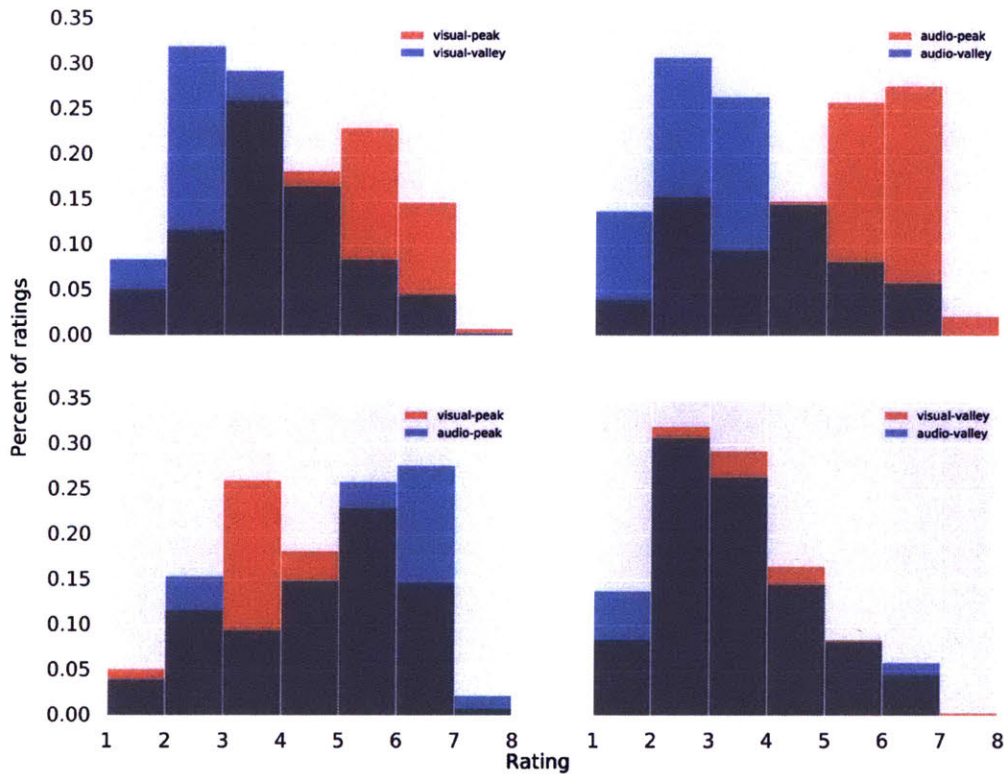


Figure 6-8: Valence Ratings: Cuts

The precision of these cuts are shown in Table 6.3, with the distribution of valence ratings shown in Figure 6-8.

Cut	Precision
Audio	0.724
Visual	0.640

(a) Audio vs. Visual

Cut	Precision
Peaks	0.596
Valleys	0.759

(b) Peaks vs. valleys

Cut	Precision
Audio-peaks	0.683
Audio-valleys	0.758
Visual-peaks	0.508
Visual-valleys	0.757

(c) Cuts

Table 6.3: Precision of clips: various cuts

We highlight and summarize the results in three broad statements:

- Audio is more precise than visual.
- Valleys are more precise than peaks.
- Audio-peaks have low precision. This is also reflected in the bottom left subplot of Figure 6-8, where the mode rating is 3 out of 7. The poor precision of audio-peaks is explored in section 6.5 and used to motivate feature engineering for the combined model in section 6.6.4.

6.5.4 Precision of audio

Next, we examine the precision of clips extracted based on the audio emotional arc, recalling that they were selected at points with varying confidence interval sizes. We examine data for our hypothesis that clips selected at points with smaller confidence intervals (defined by the standard deviation) would be more accurate.

As shown in table 6.4, smaller confidence intervals do indeed correspond with increased precision.

6.5.5 Precision by genre

Finally, we use the CMU Movie Summary Corpus [12] to tag each movie with genres. Each movie can belong to multiple genres. We then calculate precision per genre,

Stddev	Audio-peak precision	Audio-valley Precision
[0, 0.2)	4 / 4 = 1.0	81 / 88 = 0.921
[0.2, 0.4)	11 / 11 = 1.0	38 / 56 = 0.679
[0.4, 0.6)	28 / 40 = 0.7	21 / 35 = 0.6
[0.6, 0.8)	39 / 60 = 0.65	13 / 21 = 0.619
[0.8, 1.0)	43 / 68 = 0.632	8 / 23 = 0.615

Table 6.4: Precision of clips extracted from audio emotional arc: smaller confidence intervals are more precise

with a subset of results shown in table 6.5, rank-ordered by the precision of clips extracted from visual-peaks.

Genre	Overall	Visual-peak
Action	0.678	0.264
Science Fiction	0.699	0.333
Thriller	0.678	0.382
Crime Fiction	0.643	0.438
Adventure	0.726	0.443
Drama	0.660	0.520
Fantasy	0.769	0.590
Comedy	0.705	0.667
Animation	0.798	0.667
Romantic Drama	0.667	0.692
Family Film	0.760	0.722
Romance	0.678	0.757
Romantic Comedy	0.677	0.823

Table 6.5: Precision of clips: genre

While there are some differences in the overall precision, the key insight is the large variance in visual-peak precision. The relatively poor precision of visual-peaks, as noted in section 6.5.3, appears to be a product of poor precision on a number of genres.

Moreover, there is some natural grouping of the genres when listed in this order. Genres that have high visual-peak precision are lighter films falling in the romance, comedy, and family film genres. Genres that have low visual-peak precision fall in the category of action, adventure, science fiction, and thrillers.

Upon manual inspection of a few ‘incorrect’ clips, we find that visual-peak clips selected from this latter category of genres often depict, for example, outdoor battle scenes. That such scenes, which may include blood and dead bodies, would be incorrectly classified by the visual model actually makes sense considering the dataset. As the dataset was culled from publicly available images on Flickr, one would be hard pressed to find images of gore or death in the training set.

Critically, this knowledge that genre can play an important role in assessing the emotional effect of a scene is used in the next section.

6.6 Combined audio-visual model

Using the mean valence ratings as target values, we use a linear regression model to combine the audio and visual sentiment predictions.

6.6.1 Model and features

We use the Python package scikit-learn [59] to perform linear regression with polynomial interaction features of degree two. The following features, extracted from the midpoint of every clip, are used to predict the mean valence rating:

1. value of visual valence
2. (value of visual valence) - (movie’s mean visual valence)
3. (max of movie’s visual valence) - (value of visual valence)
4. (value of visual valence) - (min of movie’s visual valence)
5. peakiness* of visual valence
6. value of audio valence
7. (value of audio valence) - (movie’s mean audio valence)
8. (max of movie’s audio valence) - (value of audio valence)
9. (value of audio valence) - (min of movie’s audio valence)

10. peakiness* of audio valence
11. binned audio stddev
12. time in movie
13. movie embeddings**

where the (mean) audio/visual valence is calculated from the audio/visual arcs. An example is provided in Appendix C.

*The peakiness function $p(a, i, r)$ roughly approximates the slope and mean around the given point i for arc a (either visual or audio). r is a parameter that controls the size of the window around a_i . The function returns four values: $a_{i-1} - a_{i-r}$ (proportional to the slope left of the point), $a_{i+r} - a_{i+1}$ (proportional to the slope right of the point), $mean(a_{i-r:i-1})$ (the mean value left of the point), and $mean(a_{i+1:i+r})$ (the mean value right of the point). This covers both peaks, valleys, and inflection points. We use $r = 0.025$ for our analyses.

**The movie embedding features are described in the next section.

6.6.2 Movie embedding features

Motivated by findings regarding the impact of genre in section 6.5.5, we sought to create features that could loosely ‘summarize’ the emotional gestalt of a movie.

We use the biconcept classifier described in chapter 3 as shown in Figure 6-9. First, every frame is passed through the classifier. Next, the activations of the second to last layer are used as an embedding for that frame. We then average these embeddings across 10% of the movie, ultimately ending up with a 10×2048 matrix.

We use the dimensionality reduction algorithm TSNE [46] to visualize the 10×2048 movie embeddings. Again using the genres provided in [12], we can see if there is any correspondence between visible clusters and genre. Indeed, figure 6-10 shows a number of clusters and movies grouped loosely by genre.

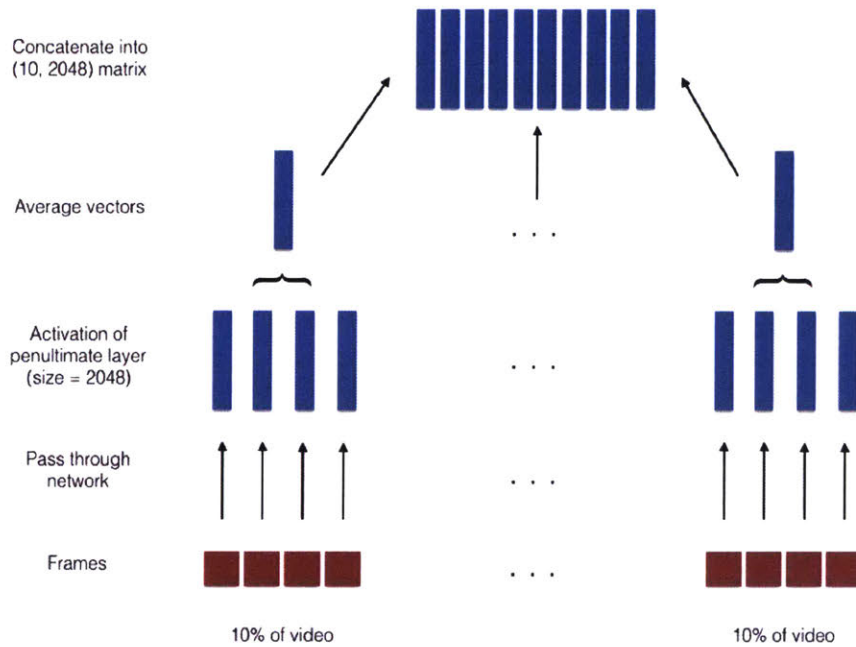


Figure 6-9: Creating movie embeddings using biconcept classifier

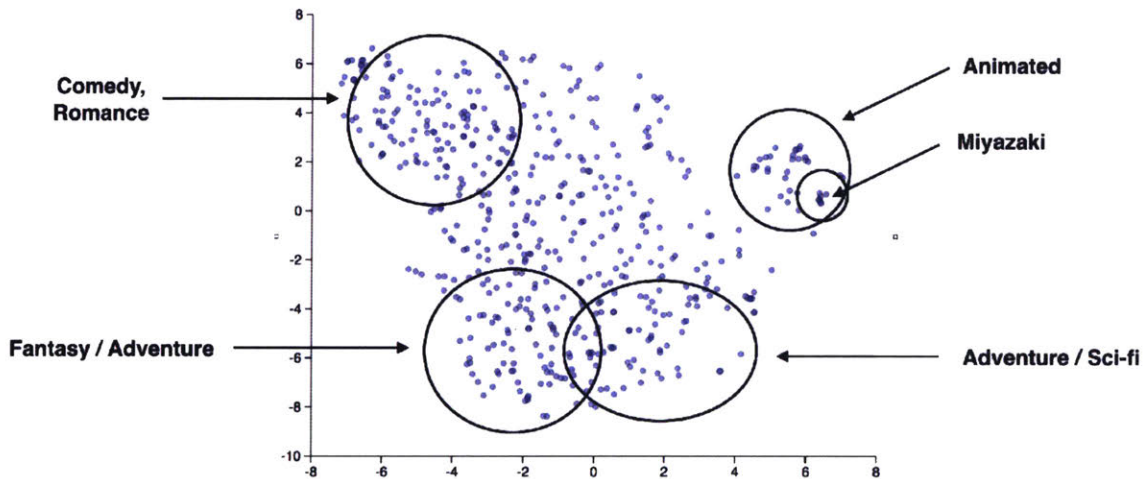


Figure 6-10: TSNE of movie embeddings, overlaid with genre

To translate these movie embeddings to features, we take the mean of each of the 10 2048-sized vectors. Thus, the final feature vector is a vector of size 10×1 .

6.6.3 Final combined precision

The performance of the final combined model is shown in Table 6.6, along with various ablations. Using all the features, we achieve a precision of 0.894.

Feature set	Overall	Aud-peak	Aud-valley	Vis-peak	Vis-valley
All features	0.894	0.940	0.884	0.872	0.886
No movie-embedding features	0.784	0.869	0.786	0.722	0.752
No ‘peakiness’ feature	0.815	0.836	0.828	0.765	0.824
Raw stddev instead of bins	0.868	0.907	0.870	0.834	0.871

Table 6.6: Precision of combined model

6.6.4 Combined arc

With this model in hand, we can create a combined emotional arc. An example is shown in Figure 6-11. Unfortunately, it appears that the combined arcs, while highly precise at certain moments, are in fact often ‘incorrect’ at a macro-scopic level. This was discovered when clustering arcs – clustering the combined arcs largely produced indistinct clusters, where most of the typical arcs were largely flat throughout the course of the movie.

This is likely due to lack of annotated data that covers the empirical joint distribution of all the features in the combined model. Over the length of the entire movie, there likely exist moments for which the model lacks representative ground truth data. For example, we don’t have clips that were extracted from flat regions in both the visual arc and the audio arc. This limitation is discussed in more detail in section 9, and we do not use the combined arc in subsequent sections.

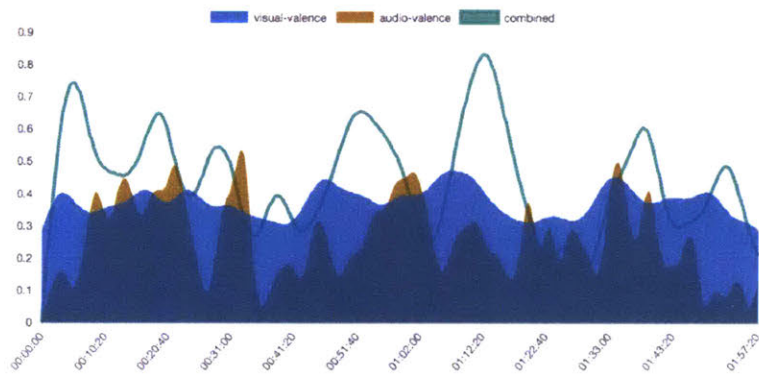


Figure 6-11: Combined emotional arc for the movie *Her*
 Frame is shown for time 1:26:29.

Chapter 7

Engagement analysis

Now armed with tools to extract emotional features and categorize emotional arcs, we can evaluate at a macro-level and return to one of our original questions – do a video’s emotional content and arc affect the degree to which people engage with it?

To tackle this question, we perform a small experiment on our Shorts Corpora in which we use a) metadata features, and b) emotional content features to predict the number of comments a video received on Vimeo. The number of comments received are shown in Figure 7-1, with the median video receiving 38 comments.

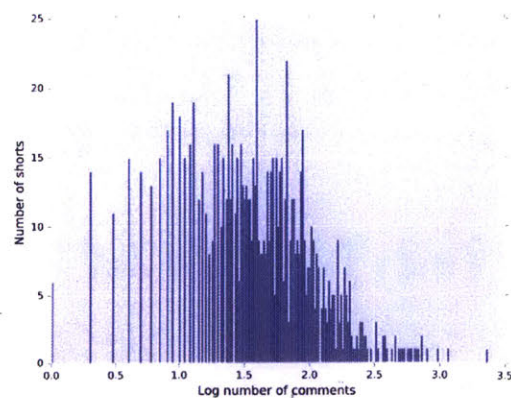


Figure 7-1: Shorts Corpora: distribution of number of comments

Video metadata features include, for example, the video length and when it was uploaded. Intuitively, we can imagine certain metadata features to be predictive variables. For instance, one might expect that longer videos receive fewer comments. Ideally, however, we would also find certain emotional features to be statistically significant predictors.

Both analyses were performed using the StatsModels Python package [65].

7.1 The effect of emotional content features

In an extensive study on what makes online content go viral, [13] and [52] demonstrated that news articles tended to spread widely if they were emotional, with positive content tending to be more viral than negative content. Would we find something similar on the stories being told and shared through video?

Feature	Coefficient	p-value
Intercept	-0.0550	0.281
<i>Duration</i>	<i>-0.0776</i>	<i>0.057</i>
Year	-0.1747	0.000
Month	-0.0554	0.115
Day	-0.0169	0.624
Hour	0.015	0.665
Author_num_comments	0.0195	0.545
<i>Visual_valence_mean</i>	<i>0.0898</i>	<i>0.090</i>
Visual_valence_std	-0.0252	0.613
<i>Audio_valence_mean</i>	<i>-0.0895</i>	<i>0.0704</i>
Audio_valence_std	0.0587	0.235
Biconcept_mean	0.0722	0.1313
Biconcept_std	-0.0562	0.176
Visual_valence_mean:Visual_valence_std	-0.0275	0.421
Audio_valence_mean:Audio_valence_std	0.093	0.018
Biconcept_mean:Biconcept_std	-0.0260	0.410

Table 7.1: Emotional content features: coefficients and p-values for model

Using a linear regression model, we investigate which features are statistically signif-

ificant predictors of the number of comments received. Features with a p-value less than 0.1 are italicized, while statistically significant features with a p-value less than 0.05 are bolded. The `biconcept_*` features are based on the biconcept-based movie embeddings described in section 6.6.2.

7.1.1 Results

Not surprisingly, longer videos tend to have fewer comments. Also, as expected, videos uploaded earlier tend to have more comments.

Intriguingly, certain emotional content features are also correlated with the number of comments. Videos that are more visually positive tend to have more comments, with a p-value of 0.090. Videos that are more negative according to the audio emotional arc, on the other hand, tend to receive more comments. Finally, the interaction feature between the mean and standard deviation of the audio valence is a statistically significant predictor – videos that have high audio valence *and* high audio variance tend to get more comments.

7.2 The effect of certain emotional arcs

Finally, the last piece of the puzzle is to assess whether certain emotional arcs themselves might be predictive of engagement.

Using the same metadata features, we run the previous analysis but with categorical cluster assignment features (which family of arcs does this movie belong to).

Following the reasoning given at the start of Chapter 5 and the description of audio given at the start of Chapter 4, we use the clusterings based on the *visual* arcs and not the audio arcs. In brief, we believe the visual arc to be a better global descriptor of a movie than the audio arc. We also don't use the combined arcs due to limitations

described in Section 6-11 and expounded upon in Chapter 9.

Nine models are created – one for each value of k in the range [2,10]. A total of three emotional arcs are statistically significant predictors, with the arcs shown in Figures 7-2, 7-3, and 7-4. The results of the models are shown in Tables 7.2, 7.3, and 7.4.

7.2.1 Results

Are any particular shapes ‘good’ or ‘bad’ in terms of audience engagement? Ultimately, we found three arcs to be statistically significant variables, each predictive of a greater number of comments. We also qualitatively examine a few prototypical films for each arc by examining the story and comments. These prototypical shorts are defined and selected by their DTW distance to the typical arc.

7.2.1.1 Good shape #1.

The arc shown in Figure 7-2 could fit the “Icarus” (rise-fall) family of stories. The model results are listed in Table 7.2. One can imagine that a story that ends so bleakly could leave quite an impact on a viewer.

Example 1 – Nowhere Line: Voices from Manus Island.¹ This short tells the story of immigrants seeking asylum in Australia, and the riot that erupted on their island detention center. An animated recreation of their journey and time there is narrated by recordings of two immigrants, Behrouz and Omar. Backgrounded by the threat of death from thirst or starvation, there is nevertheless a visual peak in the middle of the short, depicting Behrouz’s journey across the seas. At this point, there is still hope for a life better than the one he was escaping in Indonesia. This better

¹<https://vimeo.com/152158702>

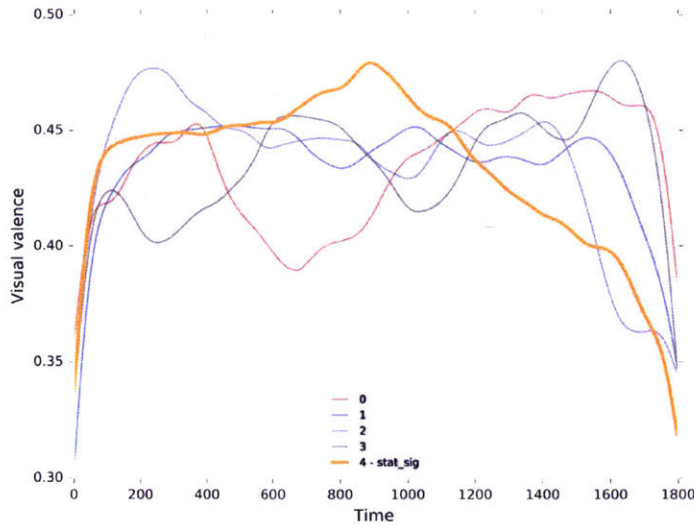


Figure 7-2: Statistically significant arc in $k = 5$ clustering

Feature	Coefficient	p-value
Duration	-0.1180	0.001
Year	-0.1720	0.000
Month	-0.0688	0.048
Day	-0.0106	0.759
Hour	0.0105	0.761
Author_num_comments	0.0053	0.878
k=5 cluster 0	-0.1574	0.203
k=5 cluster 1	-0.0729	0.213
k=5 cluster 2	0.0366	0.606
k=5 cluster 3	-0.0710	0.439
k=5 cluster 4	0.1948	0.011

Table 7.2: Emotional cluster features $k=5$: coefficients and p-values for model

life was not to be found, as he was forcibly held with others who had already been kept, in cases, for over a year.

Following the riot incited by locals and police officials, the film ends on two notes. First, it shows that 23 year-old Reza Barati was beaten and killed. Omar then states “I used to have a bright future. If I want to go back to my country, I don’t know what might happen to me. But here, I have neither safety and security, and no good life. I don’t want to pray, because I have no religion. Even those who pray, their prayers

didn't work. These things don't work on Manus Island." His friend Hamid Kehazaei is then shown to have died shortly after from a treatable infection, contracted during his time on the island.

Like nearly every short, this film receives a number of short, non-specific but highly effusive comments such as "A Really Great Work!!" or "Great job! Well done". However, this particular video also has a great number of comments highlighting the story itself and the emotional impact it had on them. We highlight some of them in the following list.

- "Superb.... so so powerful... it hits you like a wrecking ball"
- "i really feel so down and frustrated. how human could be so twisted. either way this is the best short animation story ive seen so far."
- "Haunting. Oh man. You know, I hear all the time about how hard people have it. They have no idea how bad it can really get. Impossible odds. Oh man. You've done an amazing job conveying their story. It is their animated epitaph. A testimony of a life that may not have otherwise have been heard.. All that remains of them. Man...I'm really just blown away."
- "This is absolutely beautiful, yet gut-wrenchingly awful stuff. Well done to all
- "Tragic are comments... Everyone is talking just "how great is this visual work" and not caring about story." involved. All Australians should watch."
- "Heart broken, how can anybody be so inhuman."
- "This is such a terrible matter. It feels so wrong when people say stuff in this forum praising the animation. There's so much more than the animation to this."
- "This was really heavy."
- "Really effective, you put me in their shoes man"

A number of comments emphasize the *point* of the movie – it's not the "visual work" that's important, but the story itself. This speaks broadly to movies, where the visual work should be used with purpose. We see many comments highlighting how this film builds empathy by affecting them at a deep, emotional level. We also find

further evidence of the power of a story well told, with one comment applauding a story that “may not have otherwise have been heard” and another suggesting that “all Australians should watch [this video]”.

Example 2 – Double Happy.² This short tells a coming-of-age story between four teenagers, driven by the uncomfortable group dynamics and pressures special to teenage life. The story focuses on a budding relationship between Rory and Rebecca, who nevertheless feels pressure from her friends to exclude Rory from a party. Their day culminates in an explosive scene of anger and arson when the quiet, artsy Rory is pushed to his edge, simultaneously setting a corner store and his possible relationship with Rebecca in flames.

Reflecting the idea of arc as a dynamic force, many of the comments focus on the “change of character” and the “volatile storytelling”. Others again note empathy as an element of their viewing experience.

- “Amazing, Love the change of character. Very convincing”
- “powerful. very character-driven. props to kiwi cinema”
- “The turning point for the main character felt a bit unbelievable. I don’t think theres enough time to develop a character to that point in a short structured this way, would have been nicely executed if it was longer (feature length as such). Besides that I thought it was good, nicely shot.”
- “Completely unexpected yet totally realistic. I loved it.”
- “Loved the slow burn, no pun, leading to the dramatic ending. Wonderful storytelling.”
- “This impressively captures the turbulent emotions of teen angst. I also really enjoyed how the film was scored.”
- “Mercilessly volatile storytelling.”
- “Amazingly empathetic, the rare short film that could have stood to be longer.”
- “Short filmmaking at it’s finest. Beautifully shot, and I really connected with the story.”

²<https://vimeo.com/24407168>

7.2.1.2 Good shape #2

With an even larger positive model coefficient, the arcs in the $k = 8$ and $k = 10$ clusterings are both characterized by a large, sharp peak near the end of the film, immediately followed by a steep decline. In other words, these films often end with a bang. The model results are listed in Table 7.3 and 7.4.

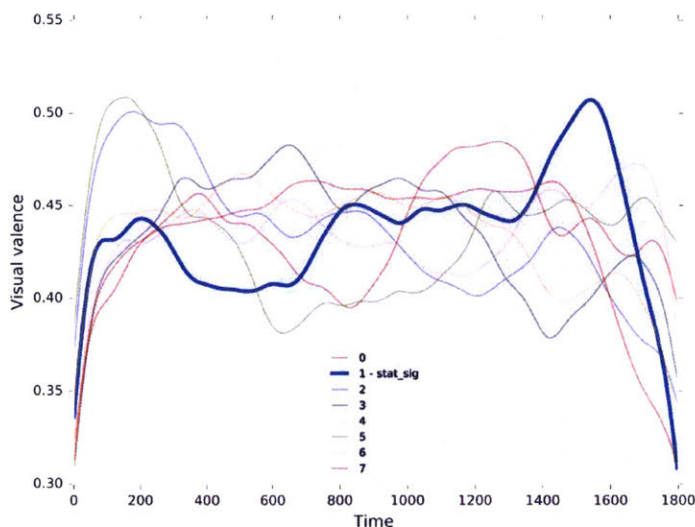


Figure 7-3: Statistically significant arc in $k = 8$ clustering

Example 1 – Benigni³ ($k = 8$, some increase before the big peak). This animated short tells the tale of a lonely xylophone player and the unusual tumor he one day grows. The tumor sprouts eyes, a mouth, and feelings to boot. A friendship emerges, and they soon do everything from tag-teaming dish washing to dressing up as Batman and Robin. The film ends as man and tumor are happily blowing bubbles out the window, only for the window to slam down and cleave off his tumor. The man is saddened but also looks on with fond remembrance.

A few users commented on the impact and suddenness of the ending. Another theme was the feeling of disgust felt while watching, with more than one user commenting on the “gross[ness]” of the film. We note this simply to show that even when negative,

³<https://vimeo.com/15596784>

strong emotions can elicit engagement from viewers.

- “Very intense!”
- “Oh man! Jumped out of my seat.”
- “..one of the saddest of sad endings.”
- “Very nice but I felt sad for the man in the film. :(heartbreaking.”
- “That was beautiful. Amazing how much emotion you were able to capture in that piece.”
- “At first I thought of being gross out, but it was a good film. I like how it reflected on loneliness remedied by a odd friendship.”
- “Gross, depressing and at the same time endearing and thought-provoking . Good Job.”
- “Saw this at Annecy and tracked you down, extremely dark subject matter and cuteness makes it all the more disturbing.”

Example 2 – ABE⁴ ($k = 10$, relatively flat before the big peak). A young woman awakens in a dim, blue room and finds herself strapped to a surgical table, her mouth sealed by duct tape. A robot walks in, questioning the nature of love and human desire. The robot states that he too was once programmed to love. But one day the humans he loved stopped reciprocating. The solution? To try to ‘fix’ the humans and make them love him again. Even though he has yet to succeed, he can’t give up, and he tries to operate once more on the tied-down woman. The film cuts to a new, brightly lit outdoor scene, where the robot stalks a new woman and expresses the wish that he could stop falling in love.

We include this short to illustrate two things. First, we are reminded by the last scene that relying on the visual alone can fail to capture the true sentiment. Second, a peak or valley can also be used not just as an emotional moment, but to create change and add a sense of motion to a film. Here, the last scene sharply contrasts the static, eerie blue of the former scene. The robot’s final words are tinged with a hint of sadness – perhaps this helped lead one user to comment “you did an amazing job

⁴<https://vimeo.com/64114843>

of making the viewer empathize with the antagonist.” But one could also argue that the last scene effectively both emphasizes the disturbing image seen in the previous minutes and adds another dimension to a film that may otherwise feel too flat.

A small number of comments reference the ending, with two noting creepiness and fear. Many others simply compliment the script and visual effects.

- “An incredibly well executed sci-fi horror short. The end was just plain eerie. The VFXs and acting were fantastic. Absolutely riveting to watch!”
- “The last scene with ABE scoping out his next ‘crush’ made my freakin’ hair stand on end.”
- “This is just chilling!”
- “awesome script and work”
- “Utterly blown away by this. Beautifully shot and wonderfully written, terrifying as it may be. :)”
- “Creepily beautiful! Great script! Well done!”
- “A lot of people are fixating on the VFX, which are admittedly awesome, but I was most impressed by the writing. Well done.”
- “Freaking WOW dude! That was brilliantly written, effects were amazing, just all around, and amazing movie! Loved it!”
- “A great complementary relationship between story and VFX. How did you get the financing for such quality of cgi?”

Feature	Coefficient	p-value
Duration	-0.1209	0.001
Year	-0.1720	0.000
<i>Month</i>	<i>-0.0664</i>	<i>0.058</i>
Day	-0.0079	0.820
Hour	0.0085	807
Author_num_comments	0.0048	0.891
k=8 cluster 0	-0.0594	0.680
k=8 cluster 1	0.3280	0.012
k=8 cluster 2	0.0852	0.434
k=8 cluster 3	0.0968	0.471
k=8 cluster 4	-0.0831	0.245
k=8 cluster 5	-0.0121	0.951
k=8 cluster 6	-0.0480	0.552
k=8 cluster 7	-0.0224	0.741

Table 7.3: Emotional cluster features k=8: coefficients and p-values for model

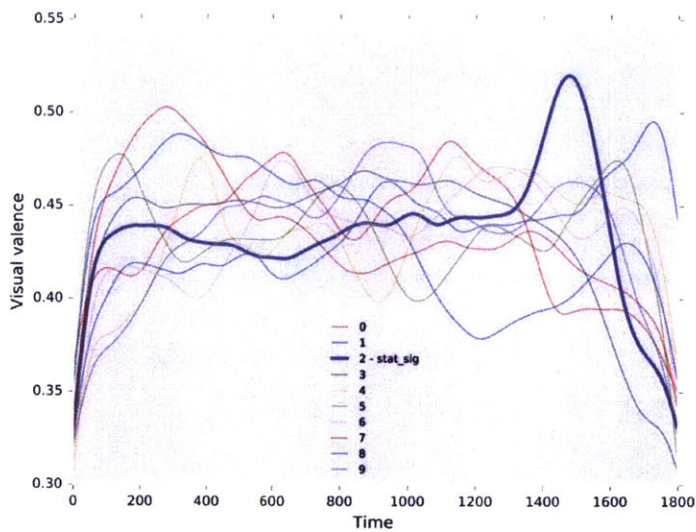


Figure 7-4: Statistically significant arc in $k = 10$ clustering

Feature	Coefficient	p-value
Duration	-0.1195	0.001
Year	-0.1771	0.000
Month	-0.0664	0.045
Day	-0.0142	0.897
Hour	0.0045	0.897
Author_num_comments	0.0110	0.753
k=10 cluster 0	-0.0110	0.932
k=10 cluster 1	-0.0962	0.423
k=10 cluster 2	0.3251	0.006
k=10 cluster 3	-0.0006	0.994
k=10 cluster 4	-0.0440	0.761
k=10 cluster 5	-0.1063	0.375
k=10 cluster 6	-0.0061	0.956
k=10 cluster 7	0.0371	0.665
k=10 cluster 8	-0.0448	0.687
k=10 cluster 9	-0.1499	0.294

Table 7.4: Emotional cluster features k=10: coefficients and p-values for model

Part III

Wrap up - where we are and beyond

Chapter 8

Related work

8.1 Emotional arcs

The research that most closely parallels this work is [64], which analyzes books from Project Gutenberg to state that “the emotional arcs of stories are dominated by six basic shapes.” Using text-based sentiment analysis, they use a singular value decomposition analysis to find a basis for emotional arcs. Looking at the bases that explain the greatest amount of variance in the emotional arcs, they are able to find typical arcs such as “rags to riches” (rise), “tragedy” (fall), “Icarus” (rise-fall), and “Cinderella” (rise-fall-rise).

In contrast to our computational approach, writers at Dramatica [5] have manually analyzed a number of books and films under the Dramatica theory of story. While they don’t explicitly create an emotional arc, many of the elements they extract are closely related, such as story outcome, emotional ‘Throughlines’, and main character growth. In fact, in response to the claims of [64], they cite their analysis of the film *Amadeus* as an example of a classic “Icarus”-style (rise-fall) emotional arc. This theory has since been used to create software that can guide storytellers in their writing.

8.2 Sentiment analysis, deep learning, and affect analysis

Research in sentiment analysis has primarily been text-based, with work ranging from short, sentence-length statements to long form articles [71, 56]. In contrast, there has been comparatively little work on images, with the Sentibank dataset [16] and [66]’s use of color distribution and SIFT-based features for visual sentiment analysis on web images being prominent examples.

Neural networks and convolutional neural networks in particular have been applied successfully in earlier research to tasks like document recognition [43]. In recent years, neural networks have seen a resurgence due in part to advances in hardware usage [42]. Deep convolutional networks have since been widely successful in both the visual and audio domain. On image-based tasks, they’ve been used for classification [42], object detection [27], and segmentation [45], among others. They’ve also been used for audio-based tasks such as automatic tagging of music [21], speech recognition [31], and speech synthesis [75].

Deep neural networks have also been previously applied to the Sentibank dataset. The authors of [20] compare their AlexNet-based results to the results of the original Sentibank paper, while [79] introduces a framework to progressively fine-tune the network by discarding noisy training examples.

Finally, we also consider our work to fall into the broad field of affect analysis. We highlight one especially relevant work, in which the authors of [49] used webcams to collect over three million facial responses from people watching over 3,000 videos. Using facial emotion recognition, this data was used in part to learn moments in videos that would trigger emotional responses.

8.3 Computational approaches to understanding story

Outside of creating emotional arcs, there exists a number of works that apply computational methods to understanding story. The M-VAD and MovieQA datasets include a combination of subtitles, scripts, and Described Video Service narrations for movies. This enables, for example, research on visual question-answering of plot [73, 69]. In the same vein of multi-modal understanding of stories, [80] looks at movies and the books they were based on, with the goal of automatically aligning scenes and passages.

There is also work centered on characters and their relationships between them. The project Movie Galaxies, for example, constructs social graphs for movies out of the interactions between characters. Elsewhere, [37] uses deep neural network-based, unsupervised topic modeling to map how character relationships evolve over the course of a novel.

Chapter 9

Limitations and future work

In this chapter, we explore the main limitations of the current system and areas for improvement. The future work is split into two broad categories. In the first, we examine *immediate extensions* that can improve or directly extend existing pieces of the system. In the second, we explore *new directions* that require greater additions to the system and expand upon ideas presented in this work.

9.1 Limitations

The main limitations center around a lack of ground truth data for movies. Though we were able to bootstrap emotional arcs by training on unrelated datasets, there are weaknesses in our system as a result. We see the problem arise in three instances:

- Lack of training data for the visual classifiers for certain commonly occurring scenes in film, leading to poorer performance in action / thriller-type genres (see section 6.5.5).
- Similarly, lack of training data covering commonly occurring sounds in films,

and thus the need for uncertainty estimates for the audio emotional arc.

- **Most critically**, lack of ground truth data for all the various types of scenes found in movies, resulting in a poorer combined emotional arc (see section 6.6.4).

Ideally, we would have many more annotated 30-second clips. Not only would we be able to train on more relevant data, we could also train a joint audio-visual model.

Other limitations include:

- A limited ability to relate the core emotional arcs to existing groupings of movies, whether that be genre, studio, release year, budget, etc.
- Finding core emotional arcs on ‘only’ a set of 509 films – this could be expanded to a larger dataset that also includes older films.
- Little interaction with filmmakers, video game designers, storytellers, and others who have may have useful domain knowledge about how this work reflects and/or could help their creative process.
- Limited understanding of what kind of scenes prompt varied emotional responses, and the degree to which they may vary depending on culture, context, and other dimensions.
- No modeling of dialogue or relationship between characters.

9.2 Future work

9.2.1 Extensions

For each extension, we attempt to note:

- Where or why it would improve the system.
- *How much it would help* the system.
- *How difficult it would be* with respect to extending the current system.

The last two points serve to inform how to prioritize work in extending the system.

Image models. We would likely see decent gains in the performance of the image classifiers by using a model architecture closer to the state of the art. Common architectures include, for example, a Residual Network [33], which uses skip connections between layers to create deeper networks, or an Inception Network [68], which uses multi-scale convolution layers. This could enable the sentiment classifier to do a better job of capturing the semantics that affect sentiment. However, we note performance should be measured not by the increases in accuracy in classification performance, but on downstream tasks regarding emotional arcs and engagement analysis. As such, it's unclear how helpful this upgrade would be. These networks may sometimes be harder to train, but in general swapping image model architectures is relatively straightforward.

Other image tasks. Similarly, we could find gains by extended image modeling through facial expression recognition and image aesthetics. For instance, the AVA dataset [54] contains labels for how aesthetically pleasing an image is. To motivate this task, we note how much image aesthetics are considered on popular services such as the photo-sharing app Instagram. In the case that this work extends to videos beyond high quality films, this could be a useful signal when analyzing short form videos that spread on social networks. Training with the AVA dataset should be straightforward.

Video modeling. We could potentially train visual models at the video-level instead of the image-level. There is some work prior work here – [38] predicted emotions on 7,701 videos from YouTube and Flickr, and [57] built a dataset of 156,219

videos by using Sentibank adjective-noun pairs as search terms on YouTube. However, we question the amount of improvement video-level predictions can provide. While frame-level predictions can be un-smooth, video classification tasks have shown that using some form of pooling across frame-level predictions is only slightly worse than video-level predictions. We believe substantial improvement would require more sophisticated modeling of the dynamics of movie scenes, backed by some understanding of how scenes are cut and shot. We further note that this is limited by the availability of high quality, large-scale datasets.

Clustering emotional arcs. While simple to use, k-medoids may not be the most effective clustering algorithm for our task. Improvements here could be especially impactful given downstream analysis on how certain emotional arcs affect engagement. We note two promising algorithms: 1) HDBSCAN, a hierarchical, non-parametric algorithm that doesn't require every point be assigned to a cluster [19], and 2) an approach designed specifically for time series that combines Gaussian Processes with Dirichlet Processes [34]. Open source implementations of both algorithms exist.

Collecting more ground truth data. All else being equal, free data is better than not for free data. We look to two examples of websites that made annotation into a playful task. If the task is enjoyable enough, we can collect data 'for free'. GIFGIF capitalizes on the popularity of gifs by presenting people with an emotion and two gifs, after which they are asked to select the gif that best matches the emotion [4]. The ESP game paired users with another random user, after which they would collaborate to label images. As images became progressively labeled, words would become "taboo" and off-limits. This would force people to come up with rich, descriptive labels [76]. We believe that watching movie clips is a relatively enjoyable annotation task and has the potential to be built into a playful website. While this may require investing some time with no guarantee that such a website would take off, the effort may be worth spending given that this lack of ground truth data is a real limitation of the current system.

Combining audio and visual emotional arcs. Here we consider ways to improve the combined emotional arc in lieu of collecting more data. First, when creating the combined arc, we might consider modeling the linear regression using a Gaussian process, which gives confidence intervals for every prediction [62]. At the very least, we'll be more aware of when the combined arc is inaccurate. In addition, in an attempt to squeeze as much as use from our limited ground truth data as possible, we might look to [30] and see how we can model the individual annotators when considering how to weight their responses. The first option is straightforward to implement, but the second requires more effort and consideration (do we even have enough data / annotators?).

Audio event data. With large portions of movies backgrounded by audio that isn't explicitly music, it could be useful to train against a dataset of commonly occurring audio events. AudioSet is one such recently released large-scale dataset, which contains an ontology of 632 audio events pulled from various moments in YouTube videos. Given that we have more relevant (relating to movies) ground truth data for visual than audio, we could also bootstrap sound models by training on movie data against the representations and labels provided by an image classifier. For example, SoundNet explores a similar idea in which they use object recognition models to learn sound representations on unlabeled video [11]. While this extension may be interesting, we caution that the effort to train this event network and integrate with the existing sentiment network may not justify the unclear gains.

Emotion prediction. Representing emotional response as a binary positive-negative scale is hardly reflective of the complex nature of emotion. We could theoretically train classifiers to predict emotion as well. For example, the adjective-noun pairs in the Sentibank dataset could be mapped to emotions using Emolex [53], which indicates for every word in its lexicon whether it's associated with one of Plutchik's eight emotions. Even better, the MVSO dataset [23] extends upon the Sentibank dataset and includes for each adjective-noun pair the *probability* that it's associated with one of the eight emotions. Outputting emotion labels could give a richer view

of movie scenes. However, we note the predictions are likely to be much noisier than binary sentiment given the imprecise nature of emotion. The signal might only be clear enough for a few emotionally charged moments in a movie. Nevertheless, after a little data setup, training on a new dataset should be straightforward.

9.2.2 New Directions

We consider four possible follow ups to this work:

1. **Character-level emotional arcs.** One could argue that the overall emotional arc of the story is really defined and constructed from the emotional arcs of the main characters. Viewers root for likable heroes and cheer the downfall of despicable villains. Plotting the emotional arcs of individual characters could also help explain the relationships between characters, with the fall of one perhaps mirrored by the rise of another.
2. **Dialogue-based emotional arcs.** Guided by crowdsourcing results in section B.2, we could also create emotional arcs based on dialogue. While the data may be sparser throughout a movie, it could at times be a more direct signal. Simple sentiment analysis could be applied on each utterance, but a more interesting approach might be to explicitly model the characters and the utterances spoken between them.
3. **From sentiment to semantics.** The two previous points suggest that to fully understand the *sentiment* of a movie, one must also understand the *semantics* of the story. We again note research on aligning books and movies [80] and question-answering in movies [69] as work that could be useful in this endeavor.
4. **Engagement analysis on social media.** While intriguing, our engagement analysis was performed against a relatively simple engagement metric – the number of comments on a website. It would be interesting to see how, if at all, emotional features affect the way videos propagate through social media

platforms like Twitter and Reddit.

Chapter 10

Conclusion

10.1 Summary of work

In this thesis, we were motivated by the idea that emotional arcs for movies both exist and matter – that every movie takes the viewer on a journey, which in turn affects how much it resonates with them. In an effort to find support for this hypothesis, we:

- constructed audio and visual emotional arcs.
- evaluated small moments within those arcs using crowdsourced annotations.
- found families of arcs in our movie datasets.

Finally, we examined whether certain emotional arcs elicit greater engagement than others, ultimately showing that this is indeed the case for a small subset of online Vimeo shorts.

10.2 Final remarks

We find our final results encouraging from a technical point of view, but not necessarily surprising from a humanist point of view. It is no secret emotions play no small part in people's lives. The history of human development is marked by increasing biological and cultural ability for emotional expression. Since the birth of mythic culture, we have seen emotional narratives as a convincing medium for explaining the world we inhabit, enforcing societal norms, and giving meaning to our existence [48].

Stories continue to have tremendous power in society today. Not only useful for entertainment, a powerful story can also activate our interests and mobilize our actions. It has been shown, for example, that emotions are the most important factor in how we make meaningful decisions in life [44].

Technological trends suggest that stories will be increasingly told in video (mobile phones) and video-like formats (virtual and augmented reality). Work that continues to push in understanding the emotional content in these mediums could shed light on the kinds of stories that connect to our large capacity for emotion.

Finally, if a movie is shown in a theater and no one is around to watch it, does it really tell a story? A deeper understanding in this area could help any storyteller, whether it be a big budget documentary director, a freelance journalist, or even a video game designer, reach and connect to a wider audience.

Appendix A

A second method for finding families of arcs

In this section, we utilize a second approach to finding typical emotional arcs by following the singular value decomposition-based approach described in [64].

A.1 Formulation

Singular value decomposition (SVD) is a widely used linear algebra method used to decompose a matrix M into the product of a unitary matrix U , a diagonal matrix Σ , and the conjugate transpose of another unitary matrix V^T , as shown in Equation A.1. In our case, each row of M is an emotional arc.

$$M = U\Sigma V^T = WV^T \tag{A.1}$$

Each row of V^T represents a basis function for the matrix M , and can thus be interpreted as basis functions for the emotional arcs. [64] refers to these as ‘modes’, where each ‘mode’ can be interpreted as a ‘core emotional arc’. The values in row i

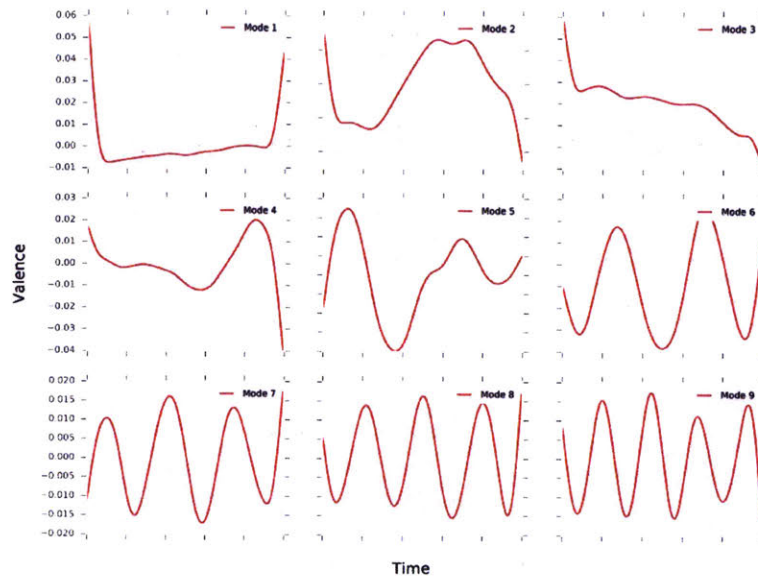


Figure A-1: Top modes for Films Corpora

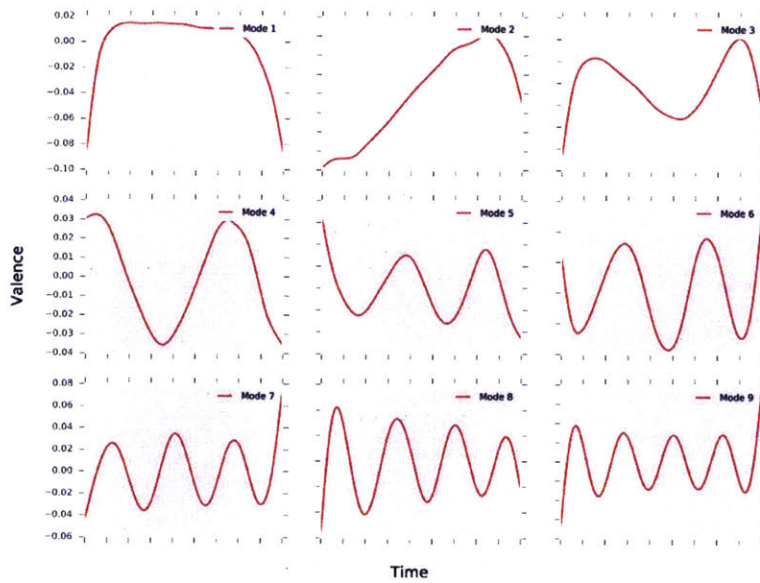


Figure A-2: Top modes for Shorts Corpora

of W represents the coefficients for weighting these basis functions in order to reconstruct the i -th emotional arc. More concretely, the contribution of mode j to the i -th emotional arc can be calculated as $W_{i,j} * V_{j,:}^T$.

A.2 Basis results

The top modes for films and shorts are shown in Figures A-1 and A-2. Note that since the coefficients can be negative, each mode has an inverse, which is simply the mode flipped across the x-axis. Again, the steep inclines and declines at the start and end, most visible in the first three modes, are artifacts of a movie's opening and closing scenes and credits.

We can also look at the films that are most similar to each mode. We illustrate this only on the Films Corpora, as we assume readers will be more familiar with the movies found in this collection. Mode 3-flipped, as shown in Figure A-3, for instance, has a slight but steady incline throughout the film, ultimately ending on a high note. That a number of the most similar films are romantic comedies makes some sense. These films tend to be lighter affairs with lower stakes and happy endings. In contrast, Mode 6, as shown in Figure A-4 experiences a series of large ups and downs. That a number of the most similar films are action-adventure epics again makes some sense.

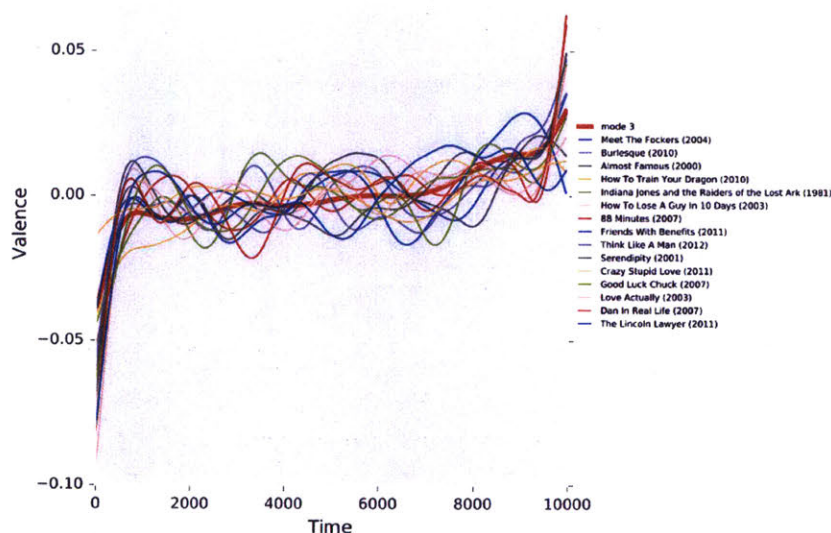


Figure A-3: Most similar films to mode 3-flipped

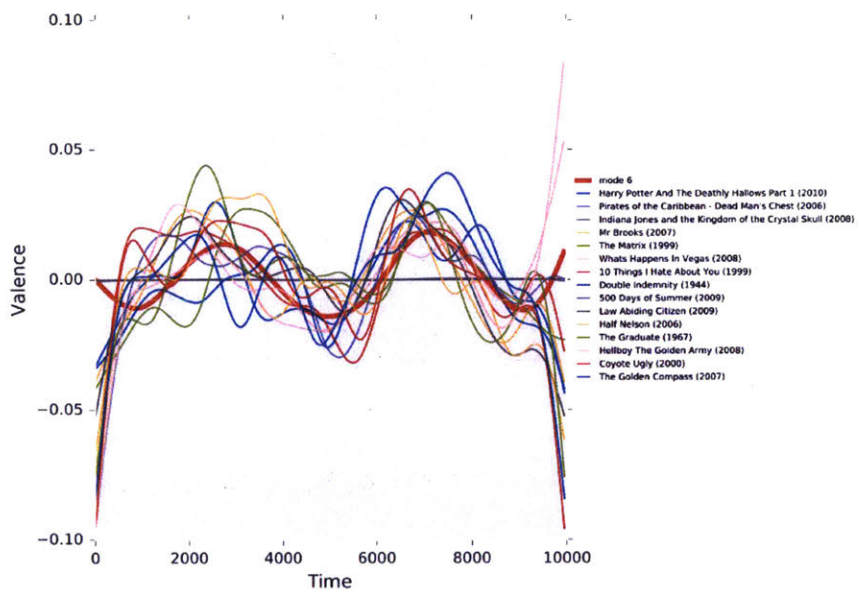


Figure A-4: Most similar films to mode 6

Appendix B

Other crowdsourcing results

B.1 Q3: Emotions in clips

Due to their more ambiguous nature, emotions were not a central part of the system. Nevertheless, the annotation results allows us to get a sense of the prevalence and relationship between emotions.

Table B.1 lists the emotions rank-ordered by their average valence rating. The average valence per emotion is calculated as the average valence rating of all clips that were labeled with the given emotion.

Emotion	Avg. valence	Avg. valence variance
Fear	2.96	2.83
Anger	3.00	2.29
Sadness	3.07	2.25
Disgust	3.26	2.60
Anticipation	3.81	2.83
Surprise	4.14	3.00
Trust	4.74	2.43
Joy	5.25	2.10
None of the above	4.01	2.08

Table B.1: Average valence of emotions

We can also look at the percent of ratings that included each emotion, broken down by positive and negative ratings. This is shown in Table B.2. In line with the valence-ordered emotions from Table B.1, *fear* occurs most frequently in negative ratings, and *joy* occurs most frequently in positive ratings.

Emotion	% of ratings	% of pos ratings	% of neg ratings
Anticipation	43.0	39.5	43.9
Joy	40.5	82.8	11.4
Sadness	35.5	16.6	51.5
Disgust	22.5	14.0	31.6
Anger	33.7	15.2	51.9
Surprise	54.8	64.4	46.8
Fear	46.7	18.1	70.6
Trust	21.4	35.1	10.7
None of the above	20.7	18.6	14.7

Table B.2: Percent of ratings that contain given emotion

However, we highlight that every ‘negative’ emotion has a not insignificant presence in positive ratings. For example, even *fear* was checked off for 18.1% of all positive ratings. The opposite also holds true for ‘positive’ emotions. We emphasize this point to reinforce the idea that an emotional response to a video clip is highly nuanced. Even a short 30-second snippet can elicit a multitude of complex emotions, some positive and others negative.

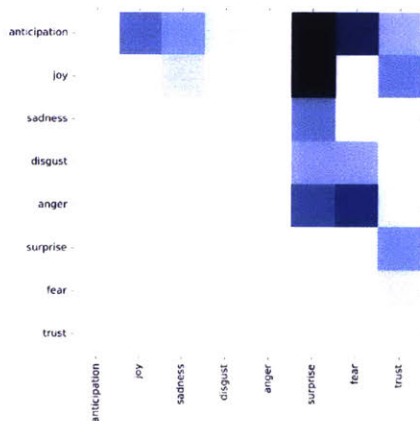


Figure B-1: Pairwise occurrences of emotions

Finally, we examine the relationships between emotions, as the emotions are non-

exclusive. Many clips were labeled with multiple emotions, and the pair-wise occurrences are visualized in a heatmap in Figure B-1. While there are some expected pairs, such as *surprise-anticipation*, we also see slightly more interesting combinations such as *sadness-surprise*. Understanding the network of emotions can lead to better modeling of emotional responses.

B.2 Q4: Audio, dialogue, visual in clips

We also analyze the responses to Q4, which asks annotators what combination of audio, dialogue, and visual input contributed to their previous answers. Annotators are allowed to select more than one answer. Understanding this is helpful for guiding how to improve the models.

We first examine the most common combinations in Table B.3. Note that the second column adds up to 100%.

Full answer	% of ratings
Audio	2.6
Dialogue	11.7
Visual	12.1
Audio + Dialogue	4.1
Audio + Visual	18.2
Dialogue + Visual	22.1
All three	29.3

Table B.3: Used to convey sentiment - full answer

Next, we examine each of the three options (audio, dialogue, visual) separately. For every rating, we calculate if a given option is present. The results are shown in Table B.4.

As expected, in both cases, visual is the most selected option. Audio is last, which affirms our suspicion that audio, while powerful, often works as a subtler emotional cue. The importance of dialogue suggests work in grounding the emotion of the scene

Answer includes	% of ratings
Audio	54.2
Dialogue	67.1
Visual	81.6

Table B.4: Used to convey sentiment - individual items

in the semantics, such as the relationship between characters.

Appendix C

Combined audio-visual model example

Imagine a video clip was extracted for annotation at 1:28:00 in Figure C-1.

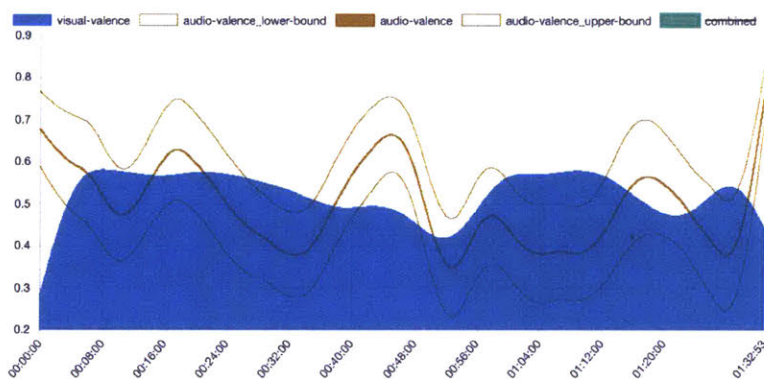


Figure C-1: Combined model: audio and visual arcs used for example

We list the features used to predict the mean valence rating given by the annotators as follows. At 1:28:00, the visual valence is 0.54, the audio valence is 0.38, and the standard deviation of the audio predictions is 0.12.

1. value of visual valence = 0.54
2. (value of visual valence) - (movie's mean visual valence) = $0.54 - 0.53 = 0.01$
3. (max of movie's visual valence) - (value of visual valence) = $0.58 - 0.54 = 0.04$

4. (value of visual valence) - (min of movie's visual valence) = $0.54 - 0.29 = 0.25$
5. peakiness* of visual valence (recall that this is four values, measuring the (proportional) slope and mean on either side of the point) = $[0.54 - 0.50 = 0.04, 0.50 - 0.54 = -0.04, 0.52, 0.52]$
6. value of audio valence = 0.38
7. (value of audio valence) - (movie's mean audio valence) = $0.39 - 0.49 = -0.10$
8. (max of movie's audio valence) - (value of audio valence) = $0.75 - 0.39 = 0.36$
9. (value of audio valence) - (min of movie's audio valence) = $0.39 - 0.35 = 0.04$
10. peakiness* of audio valence = $[0.39 - 0.44 = -0.05, 0.53 - 0.39 = 0.14, 0.41, 0.45]$
11. binned audio stddev = bin 10 (categorical variable)
12. time in movie = $1:28:00 / 1:32:53 = 0.947$
13. movie embeddings** (recall that this is a 10×1 vector) = $[0.00107333, 0.0168943, 0.0306714, 0.0330115, 0.0225832, 0.0667559, 0.0272143, 0.0361546, 0.0531328, 0.0328967]$

Bibliography

- [1] How color helps a movie tell its story. http://ideas.ted.com/how-color-helps-a-movie-tell-its-story/?utm_campaign=social&utm_medium=referral&utm_source=facebook.com&utm_content=ideas-blog&utm_term=art-design Accessed: 2017-04-27.
- [2] Kurt Vonnegut on the Shapes of Stories. <https://www.youtube.com/watch?v=oP3c1h8v2ZQ> Accessed: 2017-04-27.
- [3] Emotions. <http://changingminds.org/explanations/emotions/emotions.htm>. Accessed: 2017-04-27.
- [4] GIFGIF: Mapping the emotional language of gifs. <http://gif.gf>. Accessed: 2017-04-27.
- [5] Dramatica: the next chapter in story development. <http://dramatica.com>. Accessed: 2017-04-27.
- [6] Emotion and the film scores: An empirical approach. <http://www.e-filmmusic.de/article1.htm>. Accessed: 2017-04-27.
- [7] Introducing contributor performance levels: Crowdfower community. <http://crowdfowercommunity.tumblr.com/post/80598014542/introducing-contributor-performance-levels>. Accessed: 2017-04-20.
- [8] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

- [9] H Porter Abbott. *The Cambridge introduction to narrative*. Cambridge University Press, 2008.
- [10] Amid Amidi. *The Art of Pixar: 25th Anniversary: The Complete Color Scripts and Select Art from 25 Years of Animation*. Chronicle Books, 2015.
- [11] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *CoRR*, abs/1610.09001, 2016.
- [12] David Bamman, Brendan O’Connor, and Noah A Smith. Learning latent personas of film characters. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, page 352, 2014.
- [13] Jonah Berger and Katherine L Milkman. What makes online content viral? *Journal of marketing research*, 49(2):192–205, 2012.
- [14] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *ISMIR*, volume 2, page 10, 2011.
- [15] Christopher Booker. *The seven basic plots: Why we tell stories*. A&C Black, 2004.
- [16] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM, 2013.
- [17] Brian Boyd. *On the origin of stories*. Harvard University Press, 2009.
- [18] Joseph Campbell. *The hero with a thousand faces*, volume 17. New World Library, 2008.
- [19] Ricardo JGB Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer, 2013.
- [20] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.
- [21] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016.
- [22] Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. *arXiv preprint arXiv:1609.04243*, 2016.

- [23] Vaidehi Dalmaia, Hongyi Liu, and Shih-Fu Chang. Columbia mvso image sentiment dataset. *arXiv preprint arXiv:1611.04455*, 2016.
- [24] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.
- [25] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 80–88. Association for Computational Linguistics, 2010.
- [26] Yarın Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.
- [27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [28] Jonathan Gottschall. *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt, 2012.
- [29] Cyril Goutte, Peter Toft, Egill Rostrup, Finn Å Nielsen, and Lars Kai Hansen. On clustering fmri time series. *NeuroImage*, 9(3):298–310, 1999.
- [30] Melody Y Guan, Varun Gulshan, Andrew M Dai, and Geoffrey E Hinton. Who said what: Modeling individual labelers improves classification. *arXiv preprint arXiv:1703.08774*, 2017.
- [31] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

- [34] James Hensman, Magnus Rattray, and Neil D Lawrence. Fast nonparametric clustering of structured time-series. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):383–393, 2015.
- [35] Cisco Visual Networking Index. Cisco visual networking index: Forecast and methodology, 2010-2015. *White Paper, CISCO Systems Inc*, 9, 2011.
- [36] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [37] Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of NAACL-HLT*, pages 1534–1544, 2016.
- [38] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *AAAI*, pages 73–79, 2014.
- [39] Leonard Kaufman. Clustering by means of medoids. *Statistical data analysis based on the L1-norm and related methods*, 1987.
- [40] Eamonn Keogh. Exact indexing of dynamic time warping. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 406–417. VLDB Endowment, 2002.
- [41] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [43] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [44] Jennifer S Lerner, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. Emotion and decision making. *Annual Review of Psychology*, 66:799–823, 2015.
- [45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
- [46] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

- [47] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.
- [48] Douglas S Massey. A brief history of human society: The origin and role of emotion in social life. *American Sociological Review*, 67(1):1, 2002.
- [49] Daniel McDuff, Rana El Kaliouby, and Rosalind W Picard. Crowdsourcing facial responses to online videos. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 512–518. IEEE, 2015.
- [50] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, 2015.
- [51] Albert Mehrabian. Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies. 1980.
- [52] Katherine L Milkman and Jonah Berger. The science of sharing and the sharing of science. *Proceedings of the National Academy of Sciences*, 111(Supplement 4):13642–13649, 2014.
- [53] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. 29(3):436–465, 2013.
- [54] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE, 2012.
- [55] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632. ACM, 2005.
- [56] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.
- [57] Lei Pang and Chong-Wah Ngo. Multimodal learning with deep boltzmann machine for emotion prediction in user generated videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 619–622. ACM, 2015.
- [58] Rob Parke, Elaine Chew, and Chris Kyriakakis. Quantitative and visual analysis of the impact of music on perceived emotion of film. *Computers in Entertainment (CIE)*, 5(3):5, 2007.
- [59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

- D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [60] Robert Plutchik. 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):344–350, 2001.
- [61] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.
- [62] Carl Edward Rasmussen. *Gaussian processes for machine learning*. 2006.
- [63] Chotirat Ann Ratanamahatana and Eamonn Keogh. Everything you know about dynamic time warping is wrong. In *Third Workshop on Mining Temporal and Sequential Data*. Citeseer, 2004.
- [64] Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31, 2016.
- [65] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.
- [66] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. Analyzing and predicting sentiment of images on the social web. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 715–718. ACM, 2010.
- [67] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [68] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
- [69] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4631–4640, 2016.
- [70] Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.
- [71] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

- [72] Ronald B Tobias. *20 MASTER Plots: and how to build them*. Writer's Digest Books, 2011.
- [73] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.
- [74] Roger S Ulrich. Visual landscapes and psychological well-being. *Landscape research*, 4(1):17–23, 1979.
- [75] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*, 2016.
- [76] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.
- [77] Kurt Vonnegut. *Palm Sunday: an autobiographical collage*. Dial Press, 2009.
- [78] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.
- [79] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. *arXiv preprint arXiv:1509.06041*, 2015.
- [80] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27, 2015.