# Measuring Effectiveness of Video Advertisements

James Hahn
*Computer Science Department*
*University of Pittsburgh*
Pittsburgh, PA, USA
jrh160@pitt.edu

Dr. Adriana Kovashka
*Computer Science Department*
*University of Pittsburgh*
Pittsburgh, PA, USA
kovashka@cs.pitt.edu

*Abstract*—Advertisements are unavoidable in modern society. Times Square is notorious for its incessant display of advertisements. Its popularity is worldwide and smaller cities possess miniature versions of the display, such as Pittsburgh and its digital works in Oakland on Forbes Avenue. too much detail; abstract should be one-para and intro/motivation should be brifer; this can go in intro, not abstract Tokyo's Ginza district recently rose to popularity due to its upscale shops and constant onslaught of advertisements to pedestrians. Advertisements arise in other mediums as well. For example, they help popular streaming services, such as Spotify, Hulu, and Youtube TV gather significant streams of revenue to reduce the cost of monthly subscriptions for consumers. Ads provide an additional source of money for companies and entire industries to allocate resources toward alternative business motives. They are attractive to companies and nearly unavoidable for consumers.

One challenge for advertisers is examining a advertisement's effectiveness or usefulness in conveying a message to their targeted demographics. A company saves time and money if they know the advertisement's message effectively transfers to its audience. This challenge proves more significant in video advertisements. Rather than constructing a single, static image of content, a video advertisement possesses hundreds of frames of data with varying scenes, actors, objects, and complexity. Therefore, measuring effectiveness of video advertisements is important to impacting a billion-dollar industry. a little brifer in this para

This paper explores the combination of human-annotated features and common video processing techniques to predict effectiveness ratings of advertisements collected from Youtube. We present the first results on predicting ad effectiveness on this dataset, with binary classification accuracy of 84%.

*Index Terms*—vision, advertisements, media, video analysis

Fig. 1. Market size of advertising industry [8]

## I. INTRODUCTION

In modern society, advertisements touch nearly every aspect of a person's everyday living. Advertisements appear on television, the internet, and mobile videogames. They are also wildly popular in videos, busy city centers, radio, billboards, and posters. Behind every advertisement, there are content, market research, and ads research teams ensuring their company's advertisements achieve the highest click rate possible, converting the audience into active consumers. In general, when creating advertisements, the content and evoked emotion are heavily taken into consideration. For example, Ford wants to develop commercials with cars. If the advertisement does not possess an automobile, it will most likely be confusing and ineffective with its target audience unless the Ford logo is pla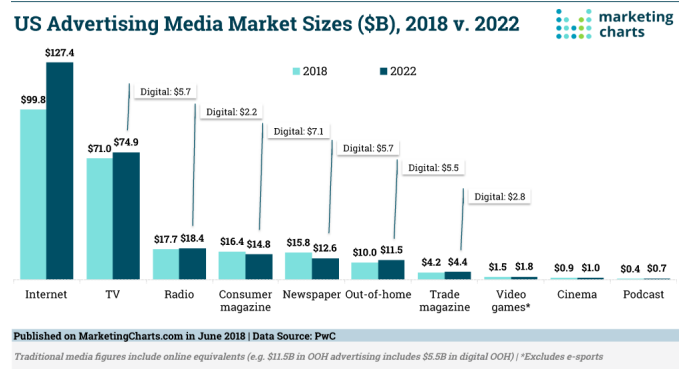ced somewhere in the video. Similarly, Ford commercials want to evoke emotions to ensure their consumers feel 'safe' or 'positive'. combine text about ford into one sentence, skip talking about objects since that seems easy/uninteresting In contrast, a company developing home security systems wants to evoke 'fear' into the audience in the sense that consumers are fearful of burglars, so they must act fast and buy a security system. Clearly, advertisements are multifaceted creations with the ability to convey symbolism, propaganda, emotions, and careful thinking.

In particular, when designing advertisements, the main goal is ensuring effectiveness. Generally, an effective ad implies a higher conversion rate for consumers buying a product. Effectiveness can be formulated by delivering a product and message in a unique, straightforward manner that stands out to consumers. The advertisements industry's market cap was over $200 billion in 2018, as seen in Figure 1 [8]. first figure should be something about our problem, not a figure from another paper– you can cite some statistics but make a different "concept figure" that allows the reader to quickly grasp what we're doing In fact, in 2017, Fox charged $5 million for a 30 second ad during Super Bowl LII, reaching more than 100 million viewers [17]. Thus, even a trivial boost in an ad's effectiveness can lead to a millions of dollars of added revenue.

combine this paragraph with the previous one, and make it briefer One rising problem with current advertisements is their failure to draw in millennials [12]. Therefore, the investigation of advertisement effectiveness is not only important

to boost revenue and prestige of a company's products, but to also engage consumers and gather their attention, especially for the next generation of consumers with buying power. Different generations respond to different stimulus, which is leading to services, such as Hulu, to implementing ad selectors [15]. These ad selectors allow viewers to select one of many displayed ads, therefore handing some of the consuming power to the consumer, rather than force feeding generalized content to large, general audiences. One research paper from 2012 investigated a similar idea to ad selectors, the video ad 'skip' feature on YouTube videos [23].

we need a paragraph that describes what we have done in this paper, at a high level, then some details about our approach/method, and a very brief overview of our results; also a statement of our scientific contribution

## II. RELATED WORK

Due to the novelty of this paper's dataset, much of the previous research with this dataset occurs in Dr. Adriana Kovashka's lab at the University of Pittsburgh. As such, these results are the first of their kind in terms of measuring advertisement effectiveness. Furthermore, this problem is two-fold: it contains the technical challenges in computer vision and the arts/media challenges in advertisements. don't talk about the dataset, don't mention names, except to cite a paper e.g. Hussain et al. [ref]; generally skip this paragraph, except to state that no prior work has studies effectiveness of ads from a computational perspective– but do a search to make sure this is true; I know for sure there are attempt to predict click-through rates in ads, which is a little different, so you should say how it's different

To our knowledge, much of the advertisements research is contained specifically in advertising journals, such as the Journal of Advertising Research from the Advertising Research Foundation [13] and the Journal of Advertising from the American Academy of Advertising [22]. As such, previous research is mostly contained in media studies, rather than the technological industry.

Additionally, within the computer vision industry, the first computer vision workshop focused on computational analysis of advertisements was hosted at CVPR18, Towards Automatic Understanding of Visual Advertisements [19]. It was the first step toward mass technological research on advertisements. don't mention workshops, events or people, mention papers

good para In terms of work related to advertisement effectiveness, previous media arts papers have focused on emotion-induced engagement [25], reduced effectiveness of sexually suggestive advertisements [2], understanding context of an ad to boost effectiveness [31], and analysis of ad content and repitition to predict effectiveness for TV commercials [27]. Other papers discussed brand recognition [21], intellectual arousal and effectiveness [28], and analysis of internet banner ad effectiveness [25]. William, in 1997, wrote a book titled Measuring Advertising Effectiveness, exploring the concept before its widespread boom into the digital age [9]. tell it in a more neutral way, don't tell a story about characters and

times– Williams explore/d... [ref] Recently, Forbes completed a survey asking readers to identify key concepts in an effective or memorable ad [11]. Responses included humor, a tagline, jingle, iconic characters, thought-provoking content, and rapport of trust and loyalty. One response surprisingly claimed hurting people creates a memorable ad. Users' answers varied and included interesting responses, but largely represented the above ideas. skip this sent This what? is clearly a difficult challenge, as "thought-provoking content" write quotes as "something" is an abstract content and is hard to convert into a computable form. [20] analyzed video advertisement effectiveness on a dataset from Akamai, but included features such as the content provider (i.e. sports, movies, news, etc.), and the position of the ad within the content (i.e. pre-roll, mid-roll, and post-roll). Additionally, they considered specific demographics of individual viewers, as well as the final result of a video's views and impressions to determine effectiveness.

the role of this para should be how what we did is different– i think this is what you're trying to do, but make it explicit: All of these works... In contrast... In this paper, effectiveness is measured by human-annotated labels, so results directly relate to whether humans believe a specific ad is effective. A given video by itself may have million of views, but it can be polarizing for its audience. For example, Justin Bieber's "Baby" music video on YouTube [3] has two billion views, but was the most disliked video on the website for years with ten million dislikes to its ten million likes. Oddly enough, YouTube's very own "YouTube Rewind 2018: Everyone Controls Rewind" video has 160m views, 2.4m likes, and 15m dislikes [29]. Clearly, views and brand are not always an accurate measure of advertisement effectiveness.

to get a new para, it's easy to just skip a line, that saves you the trouble of writing newline and indent commands– that happens automatically if you just go to a new para and skip a line

what's the role of this para? how do they relate to our work? "can be applied" is weak– if it can we should have done it, or we should say why that's insufficient Recently, there has been research on infographics, a form of modern advertisement with heavy information content [6]. Bylinskii et al. investigated textual information flow and predicted tags suitable for the content in the infographic. This same general idea can be applied to advertisements, but moreso with static images rather than video advertisements.

swap order of this and previous para since this is more related to our work Finally, in 2006, researchers investigated the use of neural networks to predict television ad effectiveness [24]. They achieved an accuracy of 99% on a 10-scale effectiveness utilizing 50 features and a dataset of 837 respondents. Therefore, [24] solved a problem similar to the one in this paper. However, there are several key differences. First, their dataset only considered television ads, while Kovashka et al. Hussain et al. considers internet, television, and general media ads [16]. Second, this dataset was collected in 2006, before

the rise of mass digital adoption. Mobile phones were barely popular around the time, resulting in less modernity of their dataset. Third, all participants were from India in a university setting. Kovashka's dataset consists of random sampling of citizens through the use of Amazon Mechanical Turk, thus giving a well-rounded background of demographics. Fourth, and perhaps most significant, all 837 participants in the survey analyzed one ad, which was for a toothpaste product. Our dataset consists of several magnitudes more topics and product types. start with that– would be nice to know how many samples they had; is it really just one? i see we need special access for this, i can look into it; i definitely wouldn't start with or even mention the 99% accuracy since it probably isn't as impressive as it sounds because of the unknown test setup– how many videos? is it really just one? if so train/test data is the same except different annotators? Therefore, this paper's research is considerably different from any prior research due to the dataset novelty, technological application, modernity, and success despite high variety of data. It provides the first modern mix of media studies with modern computer vision modern is mentioned a lot, and isn't obviously a good thing technological approaches to predict internet and television advertisement effectiveness with a diverse audience and video dataset.

## III. METHODS

brief overview of the subsections

### A. Dataset

Kovashka et. al submitted a novel dataset to CVPR 2017 collected from Amazon Mechanical Turk human annotators. Hussain et al. [ref] released a dataset... The dataset consists of two parts: 65k exact number static image advertisements and 3,477 video advertisements. Both datasets are similar with few differences in features. The static image dataset possesses labels for the topic, sentiment, question/answer statement, symbolism, strategy, and slogan. In comparison, the video dataset has labels for topic, sentiment, action/reason statement, funniness, degree of excitement, language, and effectiveness.

quotes should be 'something' not 'something' Topics are in a range of thirty-eight possibilities describing the overall theme of the advertisement, such as 'cars and automobiles', 'safety', 'shopping', or 'domestic violence'. Sentiments describe emotion evoked in the user, such as 'cheerful', 'jealous', 'disturbed', 'sad', and more, with thirty possibilities. Funniness and excitement are binary variables with value 0 indicating unfunny/unexciting and 1 implying funny/exciting. The ternary language feature takes the value 0 if the advertisement is non-English, 1 if it English, and -1 if language is unimportant to the video, such as a voiceless ad. Action/reason statements consist of a simple call to action and motivation statement combined into one. For example, one automobile commercial uses the action/reason statement, quotes "I should buy this car *because* it is pet-friendly." Every statement's action and reason are broken up with 'because'. As such, the action asked of the consumer it to buy the car

and the reason is its pet-friendly characteristic. Action/reason statements vary in complexity throughout the dataset. Finally, the goal of this research is to predict the output label, which is the 'effective' feature. Effectiveness is also a discrete value ranging from one to five.

All videos were gathered from Youtube and verified as an advertisement rather than an unrelated video. Then, human annotators on Amazon Mechanical Turk labelled all seven features for each video. Five annotators were assigned to each video to control for possible high variance in labels but were kept anonymous. Therefore, controlling for bias in work identity is unfeasible in these experiments. Additionally, the raw version of the video dataset contains all ratings for each feature for each video, while an alternative, clean version utilizes mean or mode across all five labels of each feature to compute a simplified, or 'clean', representation of that video's ratings.

### B. Data Preprocessing

The most immediate issue in terms of preprocessing is ensuring class balance. After investigation, it was discovered we discovered, active voice throughout the dataset consists of 193 samples of effectiveness 1, 261 samples of effectiveness 2, 1319 samples of effectiveness 3, 426 samples of effectiveness 4, and 1278 samples of effectiveness 5. The overall effectiveness for a given video is computed as the mode of the five ratings for the video. In general, any numeric, discrete feature with several human-annotated labels was aggregated by use of mode to better represent the video's underlying ground truth value. which data does this apply to? be explicit/clear In case of ties, the lowest effectiveness was chosen to ensure its discrete characteristic. ?? In order to normalize to ensure class balance, the class with lowest count determined the number of randomly sampled videos from each class. Therefore, 193 samples were used for each class, reducing the dataset's size to 965 and increasing difficulty of the task significantly.

make this a separate subsection called Features, and bulletize/format info about the features so it's easier to read quickly Next, to gain further insights into the factors that contribute to an ad video's effectiveness; e.g. use boldface headers to group features and introduce each group in a new paragraph or bullet, for more readability, fourteen new features were developed and computed from each video. The first five deal with the colors and visuals: average hue, median hue, average intensity, average intensity over middle 30% of video, and average intensity over middle 60% of video. Hue is classified as a 3-dimensional vector of a pixel's red, blue, and green color value. Intensity is calculated as the greyscale value of a pixel. The latter two features attempt to gauge the most captivating portions of the video. The middle 30% of a video is a window covering 30% the size of the height and width located in the center of the image; the middle 60% is computed in similar fashion. Next, average memorability across frames is computed utilizing [18]. Duration of the video is gathered from calls to the YouTube API.
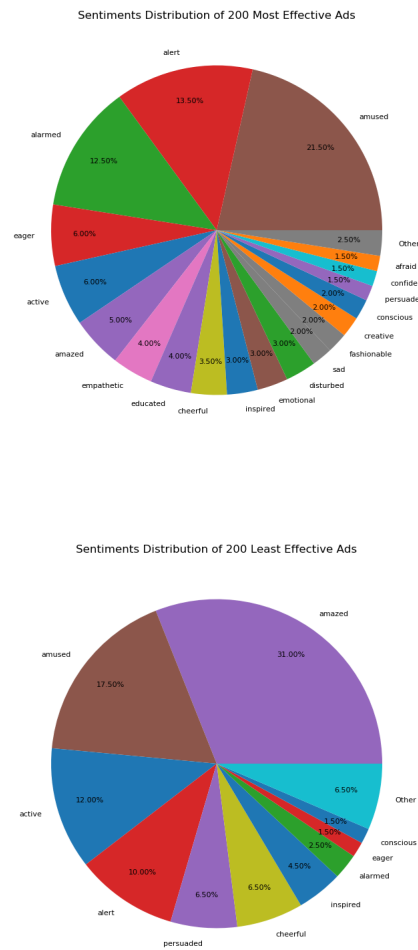
Fig. 2. Distribution of sentiments across 200 most effective and least effective advertisements
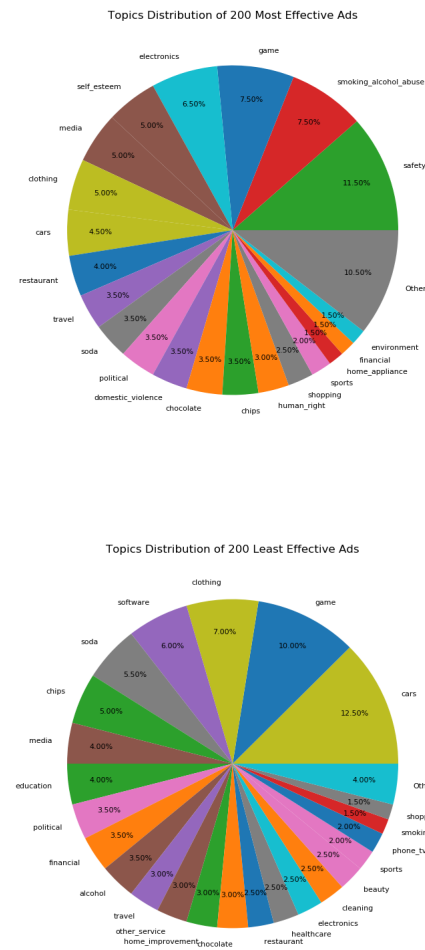


Fig. 3. Normalized distribution of topics across 200 most effective and least effective advertisements; <span style="color:red">can these be organized in descending order? it's weird to start with 0 values</span>

Text content is gathered from Google Cloud Vision's optical character recognition (OCR) API. The average duration of a video is 15 seconds at 24 frames per second, thus consisting of 360 frames. Every $60^{th}$ frame is extracted from the video, totalling 6 frames of text information per video on average. All 6 frames were passed into the API and text was extracted, resulting in four new features: average word length, keywords contained, average sentence length, and most common word. It is important to note popular text preprocessing was performed to ensure useful results, especially when counting keywords. For example, each word was stemmed (i.e. 'confused'/'confusing' become 'confuse') and stop words (i.e. 'a'/'the'/'by'/etc.) were omitted to ensure meaningfulness of features. These preprocessing techniques were handled out with the popular Python library, NLTK [4].

<span style="color:red">make sure that to the extent possible, figures show up on the same page as the place in the where they're mentioned; make sure that if you view the pdf at 100% the text is readable, and</span>

<span style="color:red">adjust any legends etc as needed</span>

Furthermore, shot boundaries were computed in addition to average optical flow of videos. Shot boundaries were measured by counting the number of scene changes throughout the video. This measures the video's quickness; higher scene changes equates to a faster video. The reason for using this metric is analyzing whether fast or slower videos translate to higher or lower effectiveness due to speed of delivery of the author's message. Additionally, optical flow is computed as the sum of the average optical flow change from frame to frame. Therefore, it is seen as a summation of vector magnitudes, representing the change in the video's content; higher optical flow is equivalent to intense content shifts. Finally, optical flow was converted into a 30-bin across the entire video. Every bin consists of the sum of vector magnitudes for that portion of frames in the video. Then, the bin was normalized via L1 norm such that all bin values sum

to one.

Ye et al.'s work investigated climax in video advertisements [34]. They gathered additional features on the video dataset including facial expressions, emotions, audio, common objects, detected climaxes, and scenes. All of these features were utilized to gain further insight into advertisement effectiveness. The audio signal per frame was averaged to represent a video's 'loudness'. Furthermore, all objects contained in the videos was gathered. Then, the probability of each object's occurrence was calculated (i.e. 1000 objects were detected, and 800 were 'person', so 'person' now has value 0.8) on the entire dataset to provide prior probabilities. Finally, each video's individual object probability distribution was calculated and each object's probability was divided by the respective prior to calculate the final distribution. For example, if the probability of a person in a specific video is 0.6, but the aforementioned prior is 0.8, then the final value in the feature vector is 0.6 / 0.8 = 0.75. This indicates a person appears 75% of the time relative to the average. In total, 786,602 objects were discovered across 80 unique classes, representing a diverse array of objects. There were 872,870 detected places/scenes, 365 of which are unique, 34,572 detected facial expressions, 8 of which were unique, and 3,659,717 detected emotions, 26 of which were unique. These three aggregated features were preprocessed in similar fashion to the objects feature (i.e. a prior was calculated to determine the probability across the entire dataset and each video's value is a ratio between its probability distribution over the prior). The audio signal (one-dimensional), object probability distribution (80-dimensional), as well as places, expressions, and emotions distributions, were used as separate feature vectors for future training. Finally, the number of climaxes per video was summed up to represent an additional feature to represent the amount of highlights in a video.

### C. Data Analysis

Simple data analysis was performed to view general trends of the dataset. Since effectiveness is the output label, correlations between each feature and the output label were computed with results demonstrated in Table 2. Correlations were performed on the entire dataset. All correlations were weak or non-existent. The features representing number of shot boundaries and number of unique annotated sentiments performed the worst. As such, fitting a linear regression line between any of the features and effectiveness will provide poor results and accuracy.
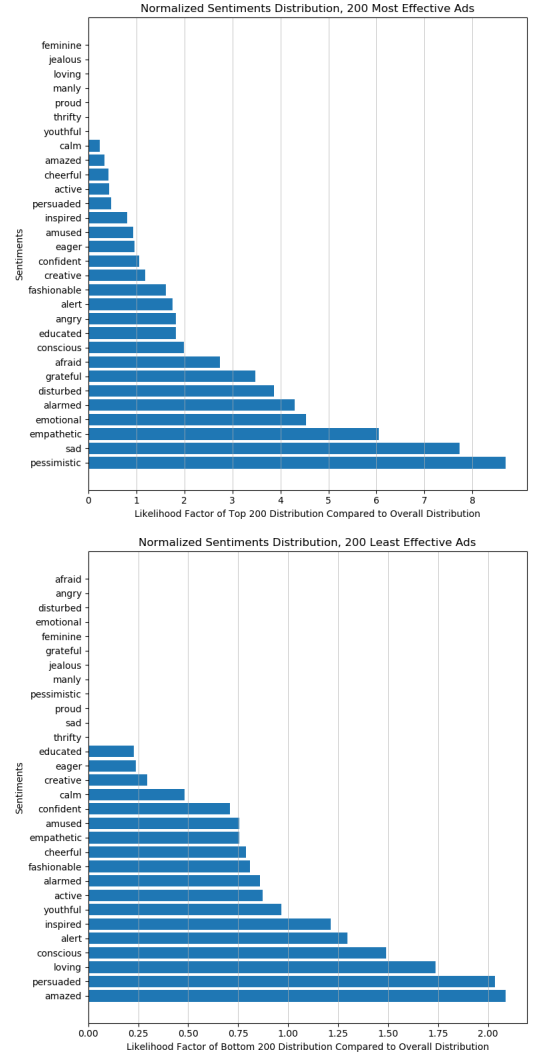


Fig. 4. Distribution of sentiments across 200 most effective and least effective advertisements

Ensuring consistency across the human annotators and their effective ratings is key to prevent outliers. Unfortunately, kappa statistics are not available for use on this dataset since anotators are anonymous and it is not possible to match one annotator's ratings to specific videos. However, the coefficient of variation, represented as $c_v = \frac{\sigma}{\mu} \times 100\%$, measures volatility of a distribution, which in this case will be the ratings for a given video, as a quality assurance check [10]. With a threshold of $c_v \leq 0.5$, there were 3114 videos (89.56%). Additionally, with thresholds of $c_v \leq 0.4$ and $c_v \leq 0.3$, there were 2932 (84.33%) and 2244 (64.54%) videos respectively. This indicates a majority of videos were under 30% variability in terms of ratings and about 20% of videos had somewhere between 30% and 40% variability. Clearly, the ratings of

| Classifier | Feature | Binary Accuracy | Four-way Accuracy | Five-way Accuracy |
|---|---|---|---|---|
| SVM | Topics | 0.5627 | 0.3178 | 0.2094 |
| SVM | Sentiments | 0.5386 | 0.2940 | 0.2435 |
| SVM | Memorability | 0.5037 | 0.2499 | 0.2072 |
| SVM | Optical Flow | 0.5060 | 0.2508 | 0.2006 |
| SVM | Cropped 30% | 0.4977 | 0.2557 | 0.1977 |
| SVM | Cropped 60% | 0.5022 | 0.2525 | 0.2142 |
| SVM | Average Hue | 0.5151 | 0.2966 | 0.2300 |
| SVM | Median Hue | 0.5102 | 0.2823 | 0.2142 |
| SVM | Duration | 0.5562 | 0.2586 | 0.2281 |
| SVM | Text Length | 0.5027 | 0.2411 | 0.2141 |
| SVM | Meaningful Words | 0.4909 | 0.2617 | 0.2158 |
| SVM | Average Word Length | 0.4966 | 0.2463 | 0.1905 |
| SVM | Word Count | 0.5064 | 0.2538 | 0.2011 |
| SVM | Sentiment Analysis | 0.4726 | 0.2480 | 0.2187 |
| SVM | Audio | 0.5246 | 0.2625 | 0.2195 |
| SVM | Objects | 0.4785 | 0.2494 | 0.2033 |
| SVM | Places | 0.5224 | 0.2676 | 0.2153 |
| SVM | Expressions | 0.4950 | 0.2366 | 0.2107 |
| SVM | Emotions | 0.4952 | 0.2574 | 0.2266 |
| SVM | Climax | 0.3589 | 0.2533 | 0.2175 |
| SVM | All Features Aggregated | 0.5669 | 0.3420 | 0.2398 |
| SVM | All Text Features Aggregated | 0.5103 | 0.2822 | 0.2269 |
| Decision Tree | Topics | 0.5720 | 0.3288 | 0.2284 |
| Decision Tree | Sentiments | 0.5464 | 0.3083 | 0.2413 |
| Logistic Regression | Exciting | 0.5595 | 0.3036 | 0.2423 |
| **Combined Ensemble** | | 0.8485 | 0.6501 | 0.5505 |
| **Baseline** | | 0.5000 | 0.2500 | 0.2000 |

**Table 1.** Accuracies of individual classifiers averaged over five simulations for binary, four-way, and five-way classification tasks.

videos were relatively consistent and no issues were found in terms of fraudulent voting. how do you know no issues? just say data is reliable... and we exclude some, if we do?

| Feature X | Correlation(X, effective) |
|---|---|
| duration | 0.207 |
| exciting | 0.181 |
| language | 0.146 |
| funny | 0.101 |
| # of detected climaxes | 0.034 |
| # of unique annotated sentiments | 0.028 |
| # of shot boundary changes | 0.026 |
| entropy of optical flow bins | -0.011 |
| avg. length of action response | -0.056 |
| audio signal | -0.071 |
| avg. length of reason response | -0.117 |

**Table 2.** Correlations between useful features and effectiveness [26] make sure you refer to the features in the same way here and in Table 1– same names and capitalizati

In search of useful indications of poor ?? effectiveness, the 200 most effective and least effective advertisements were grouped by topic and sentiment. Results can be seen in Figures 2 & 3 respectively. Keep in mind the dataset contains an uneven distribution of each topic and sentiment (e.g. 'safety' might show up three times more often than 'automobiles'). Therefore, if raw results were plotted in a pie chart, they may be skewed. For example, if videos with topic 'safety' take up 5% of the entire dataset and take up 5% of the 200 most effective ads, this is to be expected. This follows the ground truth distribution of the entire dataset. Also, Figure 4 use figure ref, don't hard-code ID represents how much more likely a given sentiment is to appear in the 200 most or least effective ads compared to the overall dataset, while Figure 5 represents the same information for topics. As supplementary analysis, topics' and sentiments' vs. effectiveness distributions are shown in Figure 6; each data point may represent more than one sample.

Finally, analysis of some individual videos in the dataset was performed by watching close to two hundred different videos, roughly 5% of the population. One video, https://www.youtube.com/watch?v=jhFqSlvbKAM, stands out since it contains Kobe Bryant and Lionel Messi, two of the most popular basketball and soccer players in the world. Also, the video itself has 146m videos on Youtube. Initial thoughts assumed it was labelled as highly effective by annotators. weird sentence, just say what conclusion was However, the mode of annotator labels was '3' with individual ratings being ['5', '3', '3', '4', '5']. As another example, https://www.youtube.com/watch?v=-usbQDfTlqE is an Atari commercial from 1982. Although the production quality is poor compared to modern advertisements, annotators listed the video as a '5' with individual ratings of ['4', '5', '5', '3', '4'], which comes out to an average of 4.2. After averaging the previous video's ratings, coming out to a '4', the second video was deemed more effective than the first despite the presence of celebrities, higher production quality, more Youtube views, and more modernity. Additionally, many videos associated with obesity, drugs, addictions, and related topics typically contained dark music, low video saturation, and sad facial expressions. A final mention is the use of positive facial expressions with product placement. For example, take https://www.youtube.com/watch?v=2Md5lPyuvsk, starring Michael Jackson. Every scene change contains a can of Pepsi next to a happy child, furthering this claim. In general, every video contains a diverse mix of features, not always
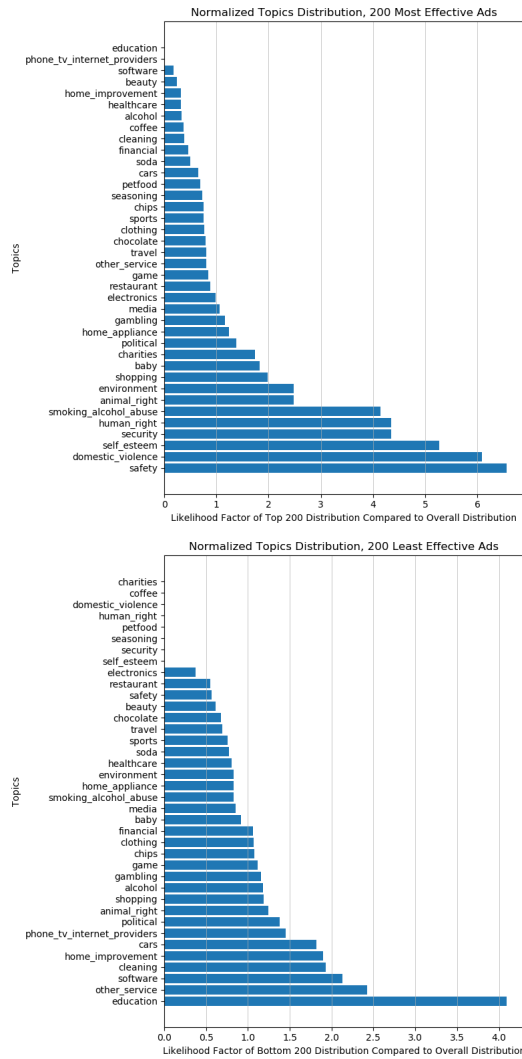
Fig. 5. Distribution of topics across 200 most effective and least effective advertisements



Fig. 6. Distribution of effectiveness ratings across each topic and sentiment

leading to the most reasonable deductive conclusion in terms of effectiveness. ?? Nonetheless, as shown, Youtube views do not always equate to higher effectiveness as mentioned earlier, and a plethora of features can be gathered from video processing to gain further insights into these advertisements' effective ratings. this is good but restructure so you're not saying "we thought X but found Y," just say what you found in Fig 6 make sure you explain what the dots are

### D. Learning

start with something positive or neutral, not a concern

The first concern of this task was the dataset's size. After preprocessing and feature gathering, 23 features spanned 594 dimensions. With a dataset of 965 total samples, a 594 dimensional feature vector is an infeasible training vector. Therefore, individual support vector machines (SVMs) [7], decision trees [32], and logistic regression [1] models were trained on each feature and a hybrid of bagging and stacking
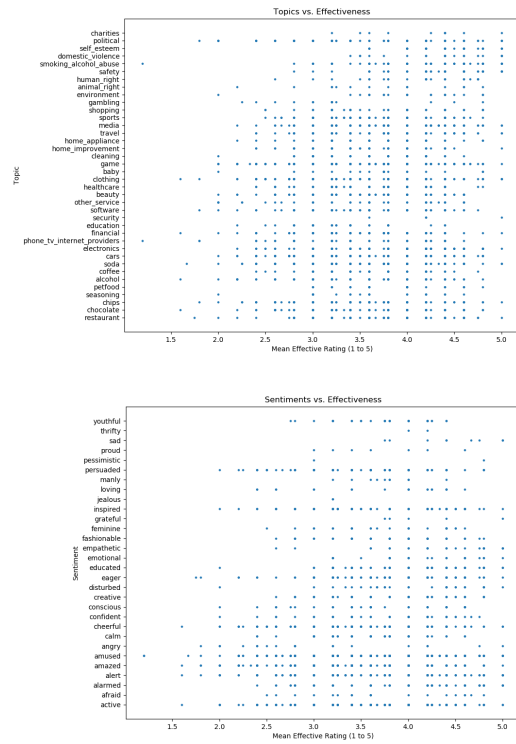
[5], two common ensemble learning techniques, was used to aggregate class predictions of the classifiers. Not only does this approach prevent overfitting, a serious concern with this data, but it also boosts confidence of the individually weak classifiers' predictions.

Specifically, each SVM was trained to achieve optimal results. Accuracies in Table 1 were the average accuracies of each SVM trained on the respective feature across five simulations. Beforehand, the dataset was randomly split for 80% training and 20% testing. Each SVM was rigorously tested in terms of the hyperparameters. For every classifier, the one-vs-rest multiclass algorithm was utilized to convert the naturally binary SVM into a multiclass classifier. This almost always produced superior results compared to one-vs-one, as well as reduced training and testing time. Different kernels worked better for different SVMs, which was discovered after rigorous testing, but most classifiers used either a polynomial or linear kernel. In Table 1, results for binary classification, four-way classification, and five-way classification across the effectiveness ratings are shown. Neural networks were avoided due to their requirement of large datasets and greater risk of overfitting.

In addition to the SVMs, a logistic regression classifier was trained on the 'exciting' feature and a decision tree was trained on the topics and sentiments features. By default, a logistic regression model in sci-kit learn utilizes one-vs-rest for multi-class classification. Probabilities of classification were not utilized as a weighting parameter in terms of accuracy or

**Algorithm 1:** Hybrid Ensemble Learning Algorithm

```
// 25 arrays for the 25 different classifiers
Topics[25][38] ← List of length 38 initialized to 0;
Sentiments[25][30] ← List of length 30 initialized to 0;
Classifiers[25] ← all 25 different classifiers to train;
for i ← 1 to 25 do
    Classifiers[i].train(trainX, trainY)
end for
for n ← 1 to N do
    for i ← 1 to 25 do
        predictedClass ← Classifiers[i].predict(trainX[n]);
        videoTopic ← video topic from 0 to 37;
        videoSentiment ← video sentiment from 0 to 30;
        if predictedClass == trainY[n] then
            Topics[i][n] += 1;
            Sentiments[i][n] += 1;
        end if
    end for
end for
for k ← 1 to 38 do
    for i ← 1 to 25 do
        Topics[i][k] /= (# videos with this topic);
        Sentiments[i][k] /= (# videos with this sent);
    end for
end for
for m ← 1 to M do
    topic ← test video topic from 0 to 37;
    sent ← test video sentiment from 0 to 30;
    c ← arg max_i [Topics[i][topic], Sentiments[i][sent]];
    predictedClass ← Classifiers[c].predict(testX[m]);
end for
```

further analysis. A decision tree was investigated for the topics and sentiments features as test runs to see if they boosted accuracy. Both tree classifiers were trained using the Gini impurity metric (default on sci-kit learn and standard in the scientific community) and new branches were created until a depth with a minimum splitting criterion was reached. Unfortunately, there was not a significant gain in information, so experimentation was stopped before using additional features. *if it didn't work well, don't describe it*

The unique hybrid ensemble learning algorithm provides the most significant boost in accuracy to predictions. *which one is that? "the" assumes it's been described?* To implement this, all individual classifiers were first trained on 80% of the dataset. Then, for each classifier, the accuracy was computed for topic and for each sentiment on the training data, resulting in 68 'bins' of accuracies. To compute accuracies on the testing dataset, the classifier with the highest accuracy for that test video's topic or sentiment *you shouldn't be using accuracy on the test set to select the classifier to use– since that means looking at test data labels!!!* was chosen to predict the output label. In this sense, a form of stacking was utilized since each classifier had a 'weight' assigned to
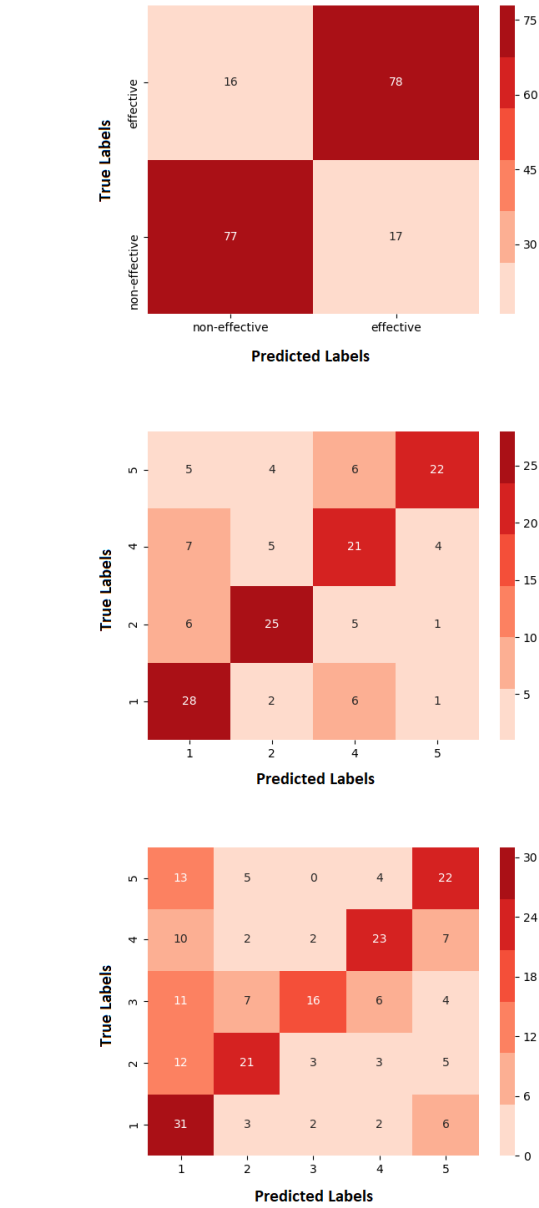


Fig. 7. Confusion Matrix of each multiclass classification task.

it (its accuracy on test video's topic or sentiment). It is also a form of bagging since each classifier votes on a class for each iteration. Only the class with the largest vote dictates the final prediction. Pseudocode of this hybrid approach is represented in Algorithm 1.

*need to have an Experiments section where you put all your results*

## IV. CONCLUSIONS

*maybe skip subsection headers here*

## A. Discussion

this part can go at the end of your experiments section
As shown, the hybrid ensemble learning algorithm produces accuracies far surpassing the baseline accuracies of 20% on five-way classification, 25% on four-way classification, and 50% on binary classification. In addition, these accuracies were achieved on small dataset of merely 965 samples. This is a significant feat since many modern machine learning and computer vision tasks require several magnitudes of samples higher in order to accomplish notable results.

Of the features, the highest in terms of accuracy was the advertisement's topic, duration, exciticement level. Also, the classifier trained on all features stood out noticeably with a 56.69% binary accuracy. On four-way classification, results were relativey the same with a few more standout features, such as the ad's sentiment and average hue. Finally, most classifiers on five-way classification performed close to the baseline, but the average hue, human emotions, all features aggregated, sentiments, and exciting classifiers produced the best results, deviating significantly from the 20% $\pm$ 1% accuracies.

In Figure 7, the three confusion matrices displaying true positives, false positives, true negatives, and false negatives are shown for all three classification tasks (five-way, four-way, binary). The binary task matrix is straightforward. For any misclassified sample, it is simply placed into the other false positive or false negative bin. Thankfully, the true positive and true negative bins contain similar numbers, so there are no significant gaps in misclassification errors. For the four-way task, most classifications were true positives. Bins with misclassified samples largely remained in the 4-6 sample range. Statistically, there are no classes with outliers in terms of misclassifications. ?? Results become a bit more concering with the five-way classification, as to be expected. don't say concerning True positive results are consistent; effectiveness 1 and 5 have highest true positive ratios. Meanwhile, effectiveness rating 3 is the most ambiguous. Clearly, with the fewest true positive ratings, many videos were rated as 2 or 4. This issue did not arise with other effectiveness ratings, other than a medium amount of samples from effectiveness 4 being predicted as effectiveness 5. Interestingly, a large number of samples from each class were predicted as effectiveness 1. This includes both classes 4 and 5, which are the most shocking. Investigation into the machine learning models, for example the support vectors, does not provide insight into the reasoning why many samples are misclassified this way.

In general, most samples were classified correctly and appeared as true positives. This is to be expected from the achieved accuracies. Furthermore, in addition to successful prediction of effectiveness ratings, statistical inference was provided for further insights into misclassifications, specific advertisement analysis, and individual feature classifier accuracies showcasing its success, rather than aggregating all features together as a blackbox.

after describing the results in expts, in the conclusion section, briefly review your method, then describe future work

## B. Future Work

Through this research, we have displayed feasibility of measuring effectiveness of advertisements through various audio, visual, and metadata features of a given video, excluding the use of Youtube comments or ratings, which are unavailable for a video before it is officially released to the public.

One future avenue of work is exploration of the static image dataset. The most significant challenge with the video dataset is the initially small sample size of 3,477 videos, which was further reduced to 965 samples after normalizing across effectiveness. The image dataset contains $\approx$ 64k images. Therefore, decreased risk of overfitting and more accurate computation of a ground-truth prior distribution for classifiers should result from the use of this dataset. It will also allow use of a validation set, many-fold cross-validation, and data modelling with higher complexity. Furthermore, the task will generally be less arduous. Videos contain more information than static images, such as duration, optical flow data, shot boundary data, climax data, and more. Despite this lack of information, static images should be more straightforward in which features to extract, resulting in higher clarity of results of the advertisement's faults.

In terms of harvesting additional information, the idea of celebrity detection was initially thought useful since popular clothing and product brands (i.e. Adidas for Lionel Messi and Nike for Lebron James) sign multi-million sponsorship deals [30]. Surely, this implies these brands value a celebrity's popularity to gain additional revenue for their brand. Also, GumGum [10], a computer vision startup based in Los Angeles, provides valuations of brands on specific sports team's jersey's, further supporting this hypothesis. However, due to monetary concerns, this feature was not explored.

A unique approach with potential to boost accuracy is training a neural network, such as a convolutional neural network, to predict the topic or sentiment of a video. In turn, this showcases how well a computer is able to represent a video's overall theme. A higher accuracy indicates an advertisement's message is successfully transferred to a human audience, increasing effectiveness.

A feat with significant capital potential is constructing a formula for effective video advertisements. For example, generating permutations of features and testing their individual effectiveness, then ranking their effectiveness. Furthermore, this ranking can become more granular for specific demographics, such as kids, adults, cultures, geographic regions, and income levels. To supplement this idea, a tool to recommend improvements to an advertisement can help boost additional revenue or outreach initiatives before the advertisement is released to the public via the internet, television, or mobile devices.

As a notable recent technological advance, Burger King has experimented with the use of artificial intelligence to generate advertisement descriptions for their products [33]. Once generated, they utilize text-to-speech software to annotate their advertisements. A potential future application is automatic

generation of video advertisements with the use of generative adversarial networks [14]. Paired with the aforementioned improvement recommendation tool, a challenge is posed to generate advertisements, provide feedback to itself, and create highly effective videos. This task is difficult due to lack of training data, high complexity of video advertisement features, and recent developments of GANs.

Finally, although the dataset provides a thorough insight into the world of advertisements, not all videos were from major brands. Therefore, a slight bias may have been introduced into the dataset. It would be interesting to investigate the comparison of these results with the results from a commercial advertisement dataset, perhaps from a major cable television network or popular streaming service.

As a result, this research has opened up several gateways to boosting how content creators generate advertisements and how consumers consume advertisements. Additionally, several avenues of research have been proposed to generate momentum within the computer vision advertisements community with new, interesting challenges.

## References

[1] A Agresti. *Categorical Data Analysis (2nd Edition)*, volume 33. 01 2002.

[2] JR Ben B. Judd and M. Wayne Alexander. On the reduced effectiveness of some sexually suggestive ads. *Journal of the Academy of Marketing Science*, 11(2):156–168, 1983.

[3] Justin Bieber. Justin bieber - baby ft. ludacris, 2010.

[4] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[5] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information science and statistics. Springer (India) Private Limited, 2013.

[6] Zoya Bylinskii, Sami Alsheikh, Spandan Madan, Adrià Recasens, Kimberli Zhong, Hanspeter Pfister, Frédo Durand, and Aude Oliva. Understanding infographics through textual and visual tag prediction. *CoRR*, abs/1709.09215, 2017.

[7] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[8] Marketing Charts. Us online and traditional media advertising outlook, 2018-2022, 2018.

[9] William D Wells. *Measuring Advertising Effectiveness*. 05 1997.

[10] B.S. Everitt and A. Skrondal. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 2010.

[11] Forbes. What makes a tv commercial memorable and effective?, 2012.

[12] Forbes. Research shows millennials don't respond to ads, 2015.

[13] American Research Foundation. Journal of advertising research, 2018.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[15] Hulu. Ad selector, 2017.

[16] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. *CoRR*, abs/1707.03067, 2017.

[17] Sports Illustrated. Super bowl LII: How much does an commercial cost?, 2018.

[18] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *International Conference on Computer Vision (ICCV)*, 2015.

[19] Adriana Kovashka. Cvpr 2018 workshop: Towards automatic understanding of visual advertisements (ADS), 2018.

[20] S. Shunmuga Krishnan and Ramesh K. Sitaraman. Understanding the effectiveness of video ads: A measurement study. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 149–162, New York, NY, USA, 2013. ACM.

[21] Hao Li and Hui-Yi Lo. Do you recognize its brand? the effectiveness of online in-stream video advertisements. *Journal of Advertising*, 44(3):208–218, 2015.

[22] American Academy of Advertising. Journal of advertising, 2018.

[23] Max Pashkevich, Sundar Dorai-Raj, Melanie Kellar, and Dan Zigmond. Empowering online advertisements by empowering viewers with the right to choose. *Journal of Advertising Research*, 52(4):451–457, 2012.

[24] V. Ramalingam, B. Palaniappan, N. Panchanatham, and S. Palanivel. Measuring advertisement effectivenessa neural network approach. *Expert Systems with Applications*, 31(1):159 – 163, 2006.

[25] Helen Robinson, Anna Wysocka, and Chris Hand. Internet advertising effectiveness. *International Journal of Advertising*, 26(4):527–541, 2007.

[26] SciPy. Pearson correlation, 2014.

[27] Surendra Singh and Catherine Cole. The effects of length, content, and repetition on television commercial effectiveness. *Journal of Marketing Research*, 30:91–104, 02 1993.

[28] Surendra N. Singh and Gilbert A. Churchill Jr. Arousal and advertising effectiveness. *Journal of Advertising*, 16(1):4–40, 1987.

[29] YouTube Spotlight. Youtube rewind 2018: Everyone controls rewind — #youtuberewind, 2018.

[30] Forbes Staff. The world's highest-paid athletes, 2018.

[31] Horst Stipp. How context can make advertising more effective. *Journal of Advertising Research*, 58(2):138–145, 2018.

[32] Paul E. Utgoff. Incremental induction of decision trees. *Machine Learning*, 4(2):161–186, Nov 1989.

[33] The Verge. Burger kings 'ai-written' ads show were still very confused about artificial intelligence, 2018.

[34] Kyle Buettner Ye, Keren and Adriana Kovashka. Story understanding in video advertisements. In *British Machine Vision Conference (BMVC)*, September 2018.