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. Tokyos 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 or to increase profit for future competitive ventures. They are attractive to companies and nearly unavoidable for consumers. One challenge for advertisers is examining the advertisements effectiveness or usefulness in conveying a message to their targeted demographics. If an advertisement is constructed, but an algorithm can predict with high accuracy whether it will be effective, and how effective in particular, businesses will save time and money. For example, a company creates an ad, but the algorithm predicts it will be ineffective, the business will have to spend more money to reimagine the ad, but potential revenue will increase since more consumers will enjoy the content after its recreation. In another scenario, a business creates an advertisement and there exist quarrels among stakeholders whether a specific portion of the ad should be changed, but the algorithm says the advertisement will be highly effective. The company will save time since they know changing the ad will cost more money and time to reshape it, while either hurting or marginally increasing effectiveness. 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 across nearly any sector of industry.**

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

I. Introduction

In progress.

II. Methods

A. Dataset

Kovashka et. al submitted a novel dataset to CVPR 2017 which was collected through use of human annotators on Amazon Mechanical Turk. The dataset consists of two parts: 65k 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.

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 an unimportant aspect of the video, such as a silent ad. Action/reason statements consist of a simple call to action and motivation statement combined into one. For example, an automobile commercial might have the action/reason statement "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 possibly high variance in labels but were kept anonymous in the dataset. 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 across all five labels of each feature to compute a simplified 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 the dataset consists of 132 samples of effectiveness 1, 331 samples of effectiveness 2, 1257 samples of effectiveness 3, 677 samples of effectiveness 4, and 1080 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. In case of ties, the lowest effectiveness was chosen to ensure its discrete characteristic. In order to normalize, the class with lowest count determined the number of randomly sampled videos from each class. Therefore, 132 samples were used for each class, reducing the dataset's size to 660 and increasing difficulty of the task significantly.

Next, to gain further insights into a video's features, fourteen new features were 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 [2]. Duration of the video is gathered from calls to the YouTube API.

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. 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 accurate results, especially when counting keywords. For example, each word was stemmed (i.e. confused/confusing become confuse) and stop words ('a'/'the'/'by'/etc.) were omitted to ensure meaningfulness of features.

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. Additionally, 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.

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 the output label were computed with results demonstrated in Table 1. These correlations were performed on the entire dataset rather than the normalized dataset since correlation is robust to change in origin or scale. All correlations were weak or non-existent with the number of shot boundaries and number of unique sentiments provided by annotators performing the worst. As such, fitting a linear regression line between any of the features and effectiveness will provide poor results and accuracy.

Feature X	**Correlation(X, effective)**
duration	0.207
exciting	0.181
language	0.146
funny	0.101
# of shot boundary changes	0.026
# of unique annotated sentiments	-0.029
avg. length of action response	-0.056
entropy of optical flow bins	-0.011
avg. length of reason response	-0.117

Table 1. Correlations between useful features and effectiveness

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 1 & 2 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.

D. Learning

In progress.

III. Conclusions

A. Discussion

In progress.

B. Future Work

In progress. In progress.

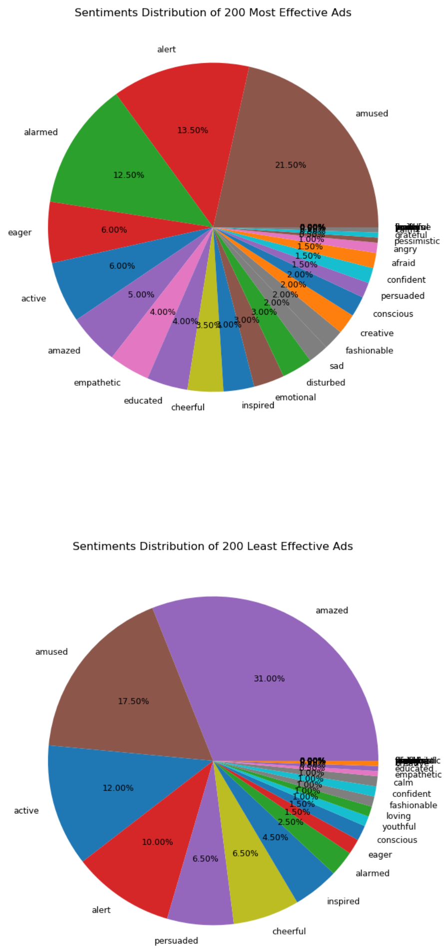

Fig. 1. Distribution of sentiments across 200 most effective and least effective advertisements

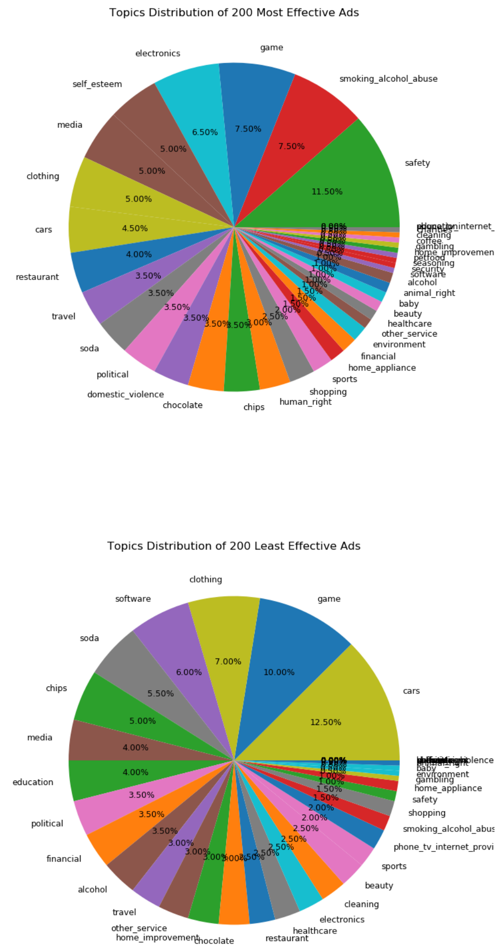

Fig. 2. Normalized distribution of topics across 200 most effective and least effective advertisements

In progress. In progress.

In progress. In progress.

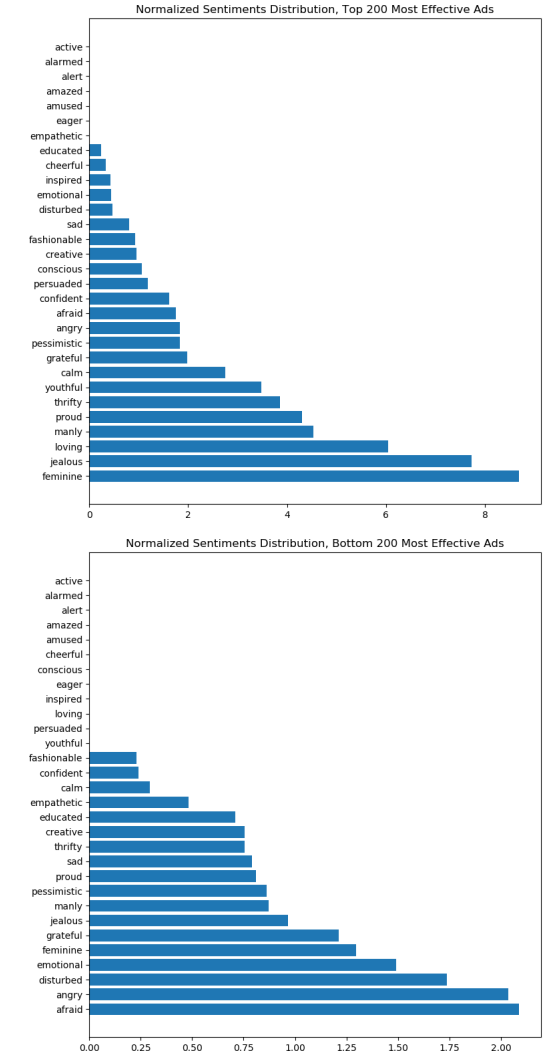

Fig. 3. Distribution of sentiments across 200 most effective and least effective advertisements

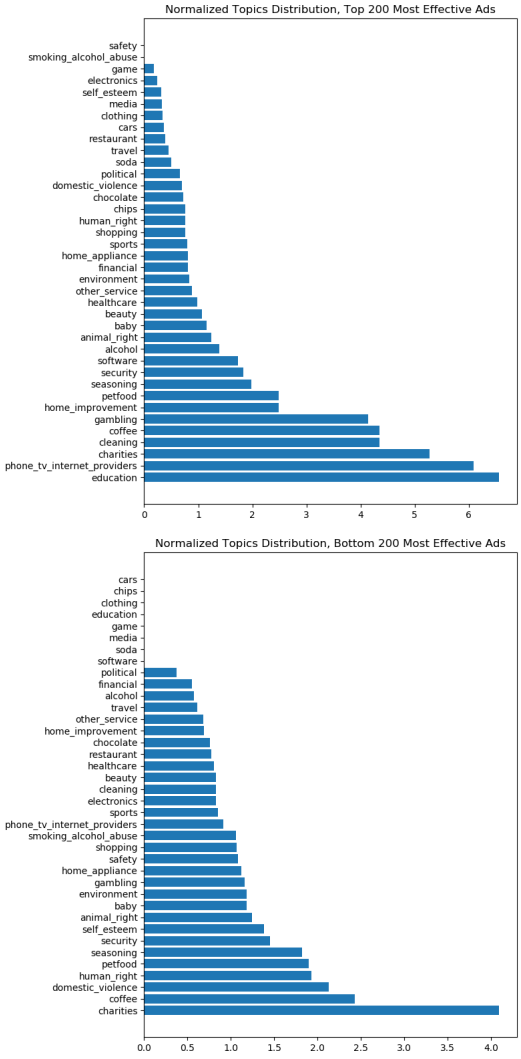

Fig. 4. Distribution of topics across 200 most effective and least effective advertisements

In progress. In progress. In progress. In progress. In progress. In progress. In progress.

REFERENCES

[1] Automatic Understanding of Image and Video Advertisements. Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, Adriana Kovashka. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017. (Spotlight)

[2] Understanding and Predicting Image Memorability at a Large Scale. A. Khosla, A. S. Raju, A. Torralba and A. Oliva. International Conference on Computer Vision (ICCV), 2015. DOI 10.1109/ICCV.2015.275

[3] In progress.

[4] In progress.

[5] In progress.

[6] In progress.

[7] In progress.

[8] In progress.

[9] In progress.

[10] In progress.

[11] In progress.

[12] In progress.

[13] In progress.

[14] In progress.

[15] In progress.

[16] In progress.