Marketing Science Institute Working Paper Series 2020
Report No. 20-147

Video Influencers: Unboxing the Mystique

Prashant Rajaram and Puneet Manchanda

Video Influencers: Unboxing the Mystique

Prashant Rajaram*
Puneet Manchanda

June 2020

This Version: December 19, 2020

* Rajaram (*prajaram@umich.edu*) is a doctoral candidate and Manchanda (*pmanchan@umich.edu*) is Isadore and Leon Winkelman Professor and Professor of Marketing, both at the Stephen M. Ross School of Business, University of Michigan, Ann Arbor, MI 48109, USA. This paper is a part of the first author's dissertation. The authors would like to thank David Jurgens, Eric Schwartz, Jun Li, Yiqi Li, Yu Song, the Marketing faculty and doctoral students at the Ross School of Business, seminar participants at Ivey Business School, University of Wisconsin-Madison, Singapore Management University, Bocconi University and the National University of Singapore for their valuable comments and feedback.

Abstract

Influencer marketing is being used increasingly as a tool to reach customers because of the growing popularity of social media stars who primarily reach their audience(s) via custom videos. Despite the rapid growth in influencer marketing, there has been little research on the design and effectiveness of influencer videos. Using publicly available data on YouTube influencer videos, we implement novel interpretable deep learning architectures, supported by transfer learning, to identify significant relationships between advertising content in videos (across text, audio, and images) and video views, interaction rates and sentiment. By avoiding ex-ante feature engineering and instead using ex-post interpretation, our approach avoids making a trade-off between interpretability and predictive ability. We filter out relationships that are affected by confounding factors unassociated with an increase in attention to video elements, thus facilitating the generation of plausible causal relationships between video elements and marketing outcomes which can be tested in the field. A key finding is that brand mentions in the first 30 seconds of a video are on average associated with a significant increase in attention to the brand but a significant decrease in sentiment expressed towards the video. We illustrate the learnings from our approach for both influencers and brands.

Keywords: *Influencer Marketing, Brand Advertising, Social Media, Interpretable Machine Learning, Deep Learning, Transfer Learning*

1. Introduction

Influencers have the capacity to shape the opinion of others in their network (Oxford Reference, 2020). They were traditionally celebrities (e.g., movie stars and athletes) who leveraged their expertise, fame and following in their activity domain to other domains. However, 95% of the influencers today, or "social media stars," are individuals who have cultivated an audience over time by making professional content that demonstrates authority and credibility (Creusy, 2016; O'Connor, 2017b). The growth in their audience(s) has been in part attributed to the fact that influencer videos are seen as "authentic" based on a perception of high source credibility. The increasing popularity of social media stars has resulted in an exponential growth of the influencer marketing industry which is expected to reach a global valuation of $10B in 2020 from $2B in 2017 (Contestabile, 2018). There are now more than 1100 influencer marketing agencies in the world that allow brands to partner with influencers to promote their products (Influencer Marketing Hub and CreatorIQ, 2020). These influencers primarily reach their audience(s) via custom videos that are available on a variety of social media platforms (e.g., YouTube, Instagram, Twitter and TikTok) (Brooks, 2020). In contrast to conventional advertising videos, influencer videos have emerged as a distinct medium (see Section 3.1 for details explaining why). Despite the rapid emergence and growth of influencer videos, there is limited research on their design and effectiveness (or indeed influencer marketing in general). Specifically, little is known about the relationship between video content and viewer reactions as well as the evolution of these videos over time.[1]

In this paper, we investigate whether the presence and nature of advertising content in videos is associated with relevant outcomes (views, interaction rates, and sentiment). There are three main challenges in carrying out these tasks. First, most data in influencer videos are unstructured. In addition, these data span different modalities – text, audio and images. This necessitates the use of state-of-the-art machine learning methods commonly referred to as deep learning. The second challenge arises from the fact that past approaches in marketing using such methods have typically made a tradeoff between predictive ability and interpretability. Specifically, such deep learning models predict marketing outcomes well out-of-sample but traditionally suffer from poor interpretability. On the other hand, deep learning models that use ex-ante handcrafted features obtain high interpretability of the captured relationships but suffer from poor predictive ability. Our "interpretable deep learning" approach handles unstructured data across multiple modalities (text, audio and images) while avoiding the need to make this trade-off. Finally, the analysis of unstructured data is computationally very demanding, leading us to use "transfer

[1] The literature on influencer marketing has primarily looked at the effect of textual content in sponsored blog posts on engagement with the post (Hughes et al., 2019) and the effect of outbound activities to other users on increasing follower base (Lanz et al., 2019).

learning." We apply our approach to publicly available influencer videos on YouTube (the platform where influencers charge the most per post[2] (Klear, 2019)).

Our approach helps us identify statistically significant relationships between marketing (brand) relevant outcomes and video elements. The significance of these relationships is supported by a significant change in attention (importance) paid by the model to these video elements. For the outcomes, we use publicly available data to develop metrics based on industry practice (Influencer Marketing Hub and CreatorIQ, 2020) and past research on visual and verbal components of conventional advertising (Mitchell, 1986). These metrics are # views, engagement (#comments / # views), popularity (# likes / # views), likeability (# likes / # dislikes) and sentiment (details are in Section 3.3). The influencer video elements we consider are text (e.g., brand names in title, captions/transcript and description), audio (e.g., speech, music, etc.), and images (e.g., brand logos, persons, clothes, etc. in thumbnails and video frames).

As noted earlier, the analysis of videos is computationally demanding, so we use a random sample of 1650 videos in order to interpret the relationship between the video elements and marketing outcomes. These videos are scraped from 33 YouTube influencers who span 11 product categories and obtain revenue from brand endorsements.[3] A concern with the use of our sample size is the possibility of "overfitting." In order to prevent that, we implement transfer learning approaches (which also have the added benefit of aiding interpretation). Transfer learning approaches, which are applied across all modalities of text, audio and image data, involve using models pre-trained (at a high monetary and computational cost) on a separate task with large amounts of data which are then fine-tuned for our different but related task. This is followed up with an ex-post interpretation step that allows identification of salient word pieces in text, moments in audio and pixels in images.

The focus on interpretation allows us to document some interesting relationships (based on a holdout sample) across all three modalities (while controlling for other variables including influencer fixed effects). First, we find that brand name inclusion, especially in the consumer electronics and video game categories, in the first 30 seconds of captions/transcript is associated with a *significant increase* in attention paid to the brand but a *significant decrease* in predicted sentiment. Second, human sounds, mainly speech (without simultaneous music), within the first 30 seconds are associated with a *significant increase* in attention, and their longer duration is associated with a *significant increase* in predicted views and likeability. Similarly, music (without simultaneous human sound) within the first 30 seconds is associated with a *significant increase* in attention. However, longer music duration is associated with a *significant decrease* in predicted engagement, popularity and likeability but a *significant increase* in

[2] An influencer with 1M–3M followers on YouTube can on average earn $125,000 per post - this is more than twice the earnings from a post on Facebook, Instagram or Twitter (O'Connor, 2017a).

[3] Our usage of this data falls within the ambit of YouTube's fair use policy (YouTube, 2020).

predicted sentiment. Third, larger pictures (of persons as well as clothes & accessories) in five equally spaced video frames (within the first 30 seconds) are associated with a *significant increase* in attention and predicted engagement. Fourth, more animal sounds in the first 30 sec of a video is associated with a *significant increase* in attention and a *significant increase* in predicted likeability. Finally, we also demonstrate that the focus on interpretability does not compromise the predictive ability of our model.

These results are relevant for multiple audiences. For academics, who may be interested in testing causal effects, our approach is able to identify a smaller subset of relationships for formal causal testing. This is done by filtering out more than 50% of relationships that are affected by confounding factors unassociated with attention (importance) paid to video elements. For practitioners, we provide a general approach to the analysis of videos used in marketing that does not rely on primary data collection. For brands, influencers and influencer agencies, our results provide an understanding of the association between video features and relevant outcomes. Influencers can iteratively refine their videos using our model and results to improve performance on an outcome of interest. Brands, on the other hand, can evaluate influencer videos to determine their impact and effectiveness at various levels of granularity (individual video elements, interactions of elements or holistic influence).

Overall, this paper makes four main contributions. First, to the best of our knowledge, it is the first paper that rigorously documents the association between advertising content in influencer videos and marketing outcomes. Second, it presents an interpretable deep learning approach that avoids making a tradeoff between interpretability and predictive ability. It not only predicts well out-of-sample but also allows interpretation and visualization of salient regions in videos across multiple data modalities – text, audio, and images. Third, it generates novel hypotheses between advertising content and a change in the outcome of interest for formal causal testing as noted above. Finally, it provides a comprehensive, data-based approach for marketers (and influencers) to assess and evaluate the quality of videos.

The remainder of the paper is organized as follows. Section 2 discusses the related literature while Section 3 describes the institutional setting and data used for analysis. Section 4 details the models for analyzing structured and unstructured data. The results are described in Section 5 while the implications of our approach and findings for practitioners (influencers and marketers) are described in Section 6. Section 7 concludes with a discussion of the limitations and directions for future research.

2. Related Literature

In this section, we review the literature on influencer marketing and unstructured data analysis (using deep learning) and describe how our work builds on it.[4]

2.1 Influencer Marketing

The nascent literature on influencer marketing has so far focused only on textual data (written text, transcripts, etc.). Hughes et al. (2019) find that high influencer expertise on sponsored blog posts is more effective in increasing comments below the blog if the advertising intent is to raise awareness versus increasing trial. However, influencer expertise does not drive an increase in likes of the sponsored post on Facebook, showing that the type of platform has a role to play in driving engagement. Zhao et al. (2019) study the audio transcript of live streamers on the gaming platform Twitch, and find that lower values of conscientiousness, openness and extraversion but higher values of neuroticism are associated with higher views. Other research, such as Lanz et al. (2019), studies network effects on a leading music platform, and finds that unknown music creators can increase their follower base by seeding other creators with less followers than creators who are influencers (with more followers). Our focus is to add to this literature by helping marketers better understand the role of text, audio and image elements in influencer videos.

2.2 Unstructured Data Analysis in Marketing via Deep Learning

The use of deep learning methods to analyze unstructured data in the marketing literature has gained increasing prominence in recent years due to its ability to capture complex non-linear relationships that help make better predictions on outcomes of interest to marketers. Marketing research on textual data has used combinations of Convolutional Neural Nets (CNNs) and Long Short-Term Memory Cells (LSTMs) to predict various outcomes including sales conversion at an online retailer (X. Liu et al., 2019), whether Amazon reviews are informative (Timoshenko & Hauser, 2019) and sentiment in restaurant reviews (Chakraborty et al., 2019). Research on image data has also used CNNs but within more complex architectures such as VGG-16 to predict image quality (Zhang et al., 2017) or classify brand images (Hartmann et al., 2020), Caffe framework to predict brand personality (L. Liu et al., 2018) and ResNet152 to predict product return rates (Dzyabura et al., 2018). Past research on both text and image data has found that deep-learning models that self-generate features have better predictive ability than those that use ex-ante hand-crafted features (Dzyabura et al., 2018; L. Liu et al., 2018; X. Liu et al., 2019). While hand-crafted features suffer from poor predictive ability, they allow interpretability of their effect on the

[4] While there is marketing literature analyzing the effectiveness of conventional advertising videos (McGranaghan et al., 2019; Teixeira et al., 2010, 2012), the difference between the two types of videos (see Section 3.1) makes it hard to use and/or build on.

outcome variable. We avoid ex-ante feature engineering of unstructured data, and instead use ex-post interpretation, so that we do not need to make a trade-off between predictive ability and interpretability.

Marketing literature has also worked with video data. Pre-trained facial expression classifiers have been used on images from video frames to infer product preference while shopping (Lu et al., 2016). Similarly, hand-crafted video features have been automatically extracted from images, audio and text of projects on the crowd funding platform Kickstarter to study their relationship with project success (Li et al., 2019). More recently, there has been research that embeds information from different data modalities using deep learning methods to create unified multi-view representations. Combinations of structured data and text have been used to predict business outcomes (Lee et al., 2018); brand logo images and textual descriptions have been combined to suggest logo features for a new brand (Dew et al., 2019); and car designs have been combined with ratings data to suggest new designs (Burnap et al., 2019). In our paper, we do not generate a modality given the other modalities, but instead focus on providing tools to improve each modality and interpreting the association between multiple modalities (text, images and audio) and our outcomes of interest.

3. Data

3.1 Institutional Setting

As noted earlier, influencer videos have emerged as a distinct marketing medium. They are quite different from conventional advertising videos[5] in at least three ways. First, these videos can (and almost always do) contain information that is unrelated to the sponsoring brand(s). This amount of information varies by type of video. On the one extreme are "integrated-advertising" videos (e.g., unboxing videos, hauls, product reviews, etc.) that feature the brand prominently throughout the video; at the other extreme are the "non-integrated-advertising" videos that feature the name of the sponsored brand only in a part of the video in the form of mini-reviews, audio shout outs, product placements or brand image displays (Mediakix, 2020). The latter type of videos includes vlogs, educational videos, gaming videos, etc. that are not directly related to the sponsoring brand(s).

Second, influencer videos are typically much longer than a standard TV commercial especially on platforms such as Instagram and YouTube.[6] By making longer videos, influencers stand to gain higher revenue from more mid-roll ad exposures. Furthermore, videos with higher expected watch time are more likely to be recommended to viewers by the YouTube recommendation algorithm (Covington et al.,

[5] Past work on characteristics of conventional advertising videos has studied their effect on ad viewing time (McGranaghan et al., 2019; Olney et al., 1991), ad attention (McGranaghan et al., 2019; Teixeira et al., 2010, 2012), ad liking / irritation (Aaker & Stayman, 1990; Pelsmacker & Van den Bergh, 1999) and purchase intent (Teixeira et al., 2014).

[6] The median duration of videos across our sample of 1650 videos is 5.3 min which is 10 times longer than the commonly used commercial duration of 30 seconds (W. Friedman, 2017).

2016). Hence, influencer video content needs to hold viewer attention for a longer duration so that the video can reach a larger audience, potentially leading to higher word of mouth and content sharing.

Third, influencer videos can be interrupted by traditional ads on YouTube. While YouTube only allows videos that are eight minutes or longer to have mid-roll ads, pre-roll ads can be a part of all influencer videos (Google, 2020c). As advertising is the primary source of revenue for influencers (Zimmerman, 2016), it is common for influencers to enable advertising on their videos, making it likely for viewers to see traditional-ad-interrupted influencer videos. Given that viewers are exposed to both influencer conveyed advertising and brand conveyed (traditional) advertising during the same viewing experience, the cognitive processing of information conveyed from each source can be quite different.

In addition to the above differences, influencer videos are also perceived to have higher source credibility (Tabor, 2020). Information about the brand is conveyed by an individual with high credibility and expertise in a related subject area, e.g., review of a beauty product coming from an influencer who has demonstrated expertise in the beauty industry.

3.2 Video Sample

We focus on 120 influencers identified by Forbes in February 2017[7] (O'Connor, 2017b). These influencers obtain revenue from brand endorsements and post mostly in English across Facebook, YouTube, Instagram and Twitter. They span 12 product categories[8] (10 influencers in each). We exclude the influencers in the Kids category as YouTube has disabled comments on most videos featuring children. Out of the remaining 110 influencers, we exclude influencers who do not have a YouTube channel. We also use the industry threshold of 1000 followers for a person to be classified an influencer (Maheshwari, 2018) and also exclude one atypical influencer with more than 100M followers. Furthermore, we short-list those influencers who have at least 50 videos so that we can capture sufficient variation in their activity, which leaves us with a pool of 73 influencers. From this pool, we randomly choose 3 influencers per category, which gives a total of 33 influencers[9] and a master list of 32,246 videos, whose title and posting time were scraped using the YouTube Data API v3 in October 2019. In addition, we also record the subscriber count for each channel at the time of scraping. From this pool of 33 influencers, we randomly choose 50 public videos for each influencer so that we have a balanced sample of 1650 videos that is feasible to analyze. Excluding videos in which either likes, dislikes or

[7] The criteria used by Forbes to identify these influencers include total reach, propensity for virality, level of engagement, endorsements, and related offline business.

[8] The 12 product categories are Beauty, Entertainment, Fashion, Fitness, Food, Gaming, Home, Kids, Parenting, Pets, Tech & Business, and Travel.

[9] Three of the randomly chosen influencers had comments disabled on more than 95% of their videos, and hence three other random influencers were chosen in their place from the respective category.

comments were disabled by the influencer(s) leaves us with 1620 videos (all scraped in November 2019). Table 1 shows the specific data scraped.

Structured Data	Metrics	Number of views (from time of posting to time of scraping)
		Number of comments (from time of posting to time of scraping)
		Number of likes (from time of posting to time of scraping)
		Number of dislikes (from time of posting to time of scraping)
	Length	Video Length (min)
	Tags	Tags associated with each video (see Google (2020a) for details)
	Playlist	Number of playlists the video is a part of
		Position of video in each playlist
		Number of videos on all the playlists the video is a part of
	Time	Time of posting video
Unstructured Data	Text	Title
		Description
		Captions (if present)
		Comments (Top 25 as per YouTube's proprietary algorithm) with replies
	Audio	Audio file
	Images	Thumbnail
		Video file

Table 1: Scraped data for videos

3.3 Outcome Variables

The top three ways of measuring influencer marketing success in the industry are conversions, interaction rates and impressions (Influencer Marketing Hub and CreatorIQ, 2020). Unfortunately, conversion data are not publicly available. We capture the remaining two (sets of) variables and in addition also capture sentiment.

(1) Impressions (Views)

Views are important not only to brands, but also to influencers. Higher views help brands increase exposure levels of their influencer marketing campaign, and help influencers earn more revenue equal to a 55% share of ad CPM[10] on YouTube (Rosenberg, 2018). Furthermore, an increase in views is correlated with an increase in channel subscribers,[11] and higher subscriber count allows the influencer to earn higher CPM rates (Influencer Marketing Hub, 2018) as well as to ex-ante charge higher for a brand collaboration (Klear, 2019; O'Connor, 2017a).

[10] Median ad CPM rates on YouTube are $9.88, and form the primary source of revenue for YouTube influencers (Lambert, 2018; Zimmerman, 2016)

[11] Total views for all videos of an influencer channel are highly correlated with subscriber count for the channel across the 33 influencers in our sample, $\rho = 0.91$.

There are a few different ways in which public view counts are incremented on YouTube. First, watching a complete pre-roll ad that is 11 to 30 seconds long OR watching at least 30 seconds of a pre-roll ad that is longer than 30 seconds OR interacting with a pre-roll ad (Google, 2020b). Second, if a pre-roll ad is skipped OR there is no pre-roll ad OR the complete pre-roll ad is smaller than 11 seconds, then watching at least 30 seconds of the video (or the full video if it has a shorter duration) has been historically documented to be the minimum requirement for public view counts to increase (Parsons, 2017).

On average, only 15% of viewers have been typically found to watch 30 seconds of a YouTube pre-roll ad (Influencer Marketing Hub, 2018). Hence, it is likely that most view counts are incremented because of viewing the first 30 seconds of video content. As views are exponentially distributed, we show the distribution of the log of views across our sample of 1620 videos in Figure 1. The distribution is approximately normal and ranges from 3.71 to 17.57 with a median of 11.85 (or 140,000 views).

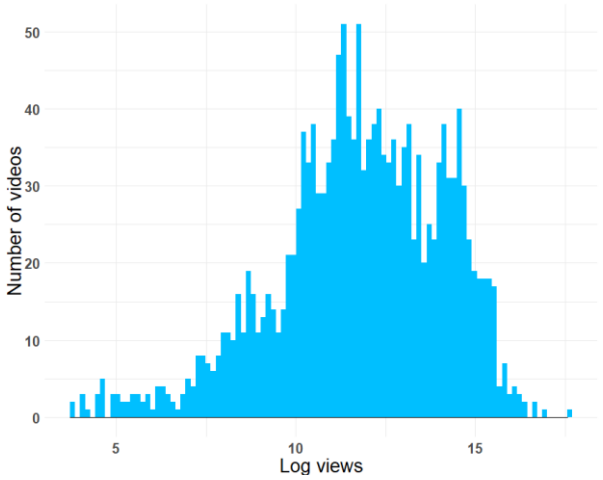

Figure 1 – Distribution of log view count

(2) Interaction Rates

Brands care more about interaction rates than views to not only ex-ante decide on a collaboration but also to ex-post measure campaign success (Influencer Marketing Hub and CreatorIQ, 2020). Hence, in addition to using impressions (views) as an outcome of interest, we develop three measures of interaction rates that are captured in publicly available data: (a) engagement = (# comments / # views), (b) popularity = (# likes / # views), and (c) likeability = (# likes / # dislikes). While measuring number of comments and likes is common practice in industry and academia (Dawley, 2017; Hughes et al., 2019), we scale each measure by (number of) views to develop unique measures that are not highly correlated with views,[12] and hence can be used to compare interaction rates for videos with different levels of views. The third metric,

[12] Across 1620 videos spanning 33 influencers, there is a high correlation between log views and log (comments+1) at 0.91, between log views and log (likes+1) at 0.95 and between log views and log (dislikes+1) at 0.92 (we add 1 to avoid computation of log(0)).

(# likes / # dislikes), is unique to YouTube because YouTube is the only major influencer platform in the US which publicly displays number of dislikes to the content.[13] As the three interaction rates are also exponentially distributed, we take their natural log, and add 1 to avoid computation of log(0) or log(∞): (a) log engagement = log $\left(\frac{\text{comments}+1}{\text{views}}\right)$, (b) log popularity = log $\left(\frac{\text{likes}+1}{\text{views}}\right)$, and (c) log likeability = log $\left(\frac{\text{likes}+1}{\text{dislikes}+1}\right)$. The distribution of the log of the interaction rates for the 1620 videos is shown in Figure 2a, 2b and 2c. The distribution of all three interaction rates is approximately normal. Log engagement has a median of – 6.21 (or 19 comments per 10K views), log popularity has a median of – 3.81 (or 220 likes per 10K views) while log likeability has a median of 3.99 (or approximately 54 likes per dislike).

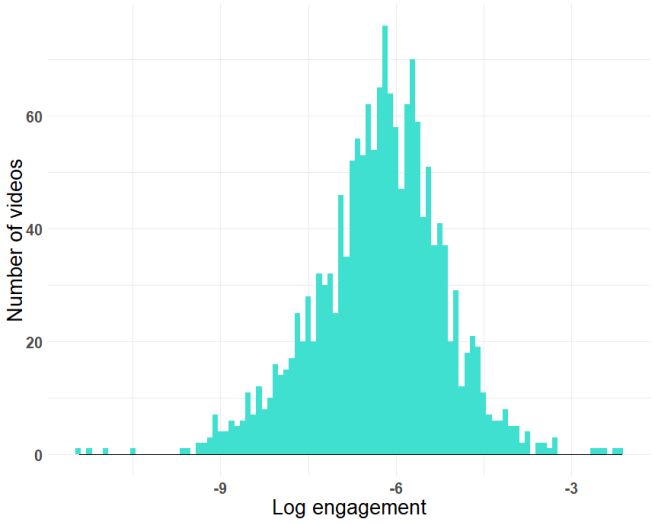

Figure 2a – Distribution of Log Engagement

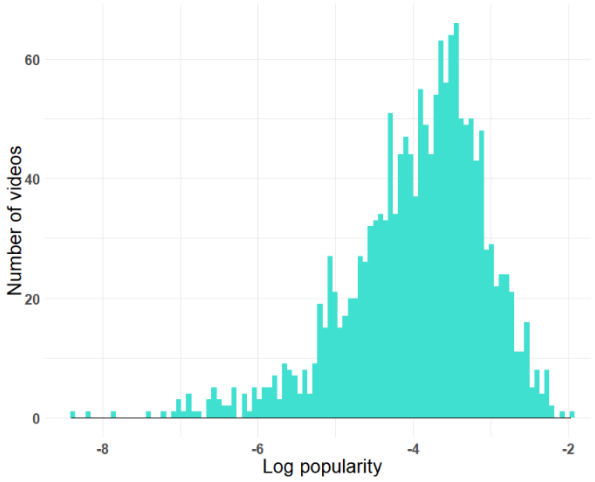

Figure 2b – Distribution of Log Popularity

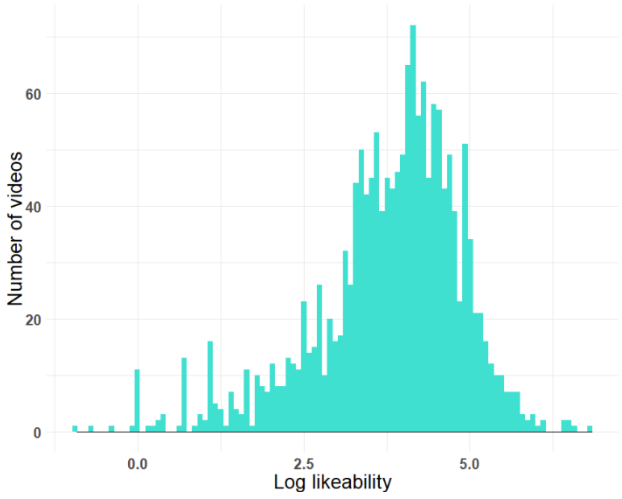

Figure 2c – Distribution of Log Likeability

[13] Other influencer platforms either do not allow dislikes to content or only allow content to be marked as 'not interesting' which is not publicly displayed.

(3) Sentiment

Past work has found that the visual and verbal components of advertising can have an effect on attitude towards the ad which in turn can have a direct effect on overall brand attitude, including attitude towards purchasing and using the product (Mitchell, 1986). Hence, it is likely that brands would benefit from understanding viewer attitude towards the video as it acts as a proxy for sales. We capture attitude towards a video by measuring the average sentiment expressed in the Top 25 comments below a video using Google's Natural Language API. Comments below a YouTube video are by default sorted as 'Top comments' and not 'Newest first,' using YouTube's proprietary ranking algorithm.[14] Note that we do not measure the sentiment in the replies to each of the Top 25 comments because sentiment expressed in the reply is likely to be sentiment towards the comment and not sentiment towards the video.

The Natural Language API by Google is pre-trained on a large document corpus, supports 10 languages, and is known to perform well in sentiment analysis on textual data (including emojis) in general use cases (Hopf, 2020). For comments made in a language not supported by the API, we use the Google Translation API to first translate the comment to English, and then find its sentiment. The sentiment provided is a score from -1 to $+1$ (with increments of 0.1), where -1 is very negative, 0 is neutral and $+1$ is very positive. We calculate the sentiment of each comment below a video for a maximum of Top 25 comments, and then find the average sentiment score.[15]

The distribution of sentiment scores for the 1620 videos is shown in Figure 3. It ranges from -0.9 to 0.9 with a median of 0.34, which we use as a cut-off to divide sentiment in the videos into two buckets – "positive" and "not positive (neutral or negative)." The large peak at 0 is because of 71 videos where viewers do not post any comments (even though comment posting has not been disabled by the influencer). We assume that if viewers choose to not post comments below a video, then the sentiment towards the video is neutral (0).

[14] Higher ranked comments (lower magnitude) have been empirically observed to be positively correlated with like/dislike ratio of comment, like/dislike ratio of commenter, number of replies to the comment and time since comment was posted (Dixon & Baig, 2019). Moreover, a tabulation shows that 99% of comments are made by viewers and not the influencer (who owns the channel) and hence we do not separate the two.

[15] As a robustness check, we use Top 50 and Top 100 comments for a random sample of 66 videos (2 videos per influencer) and also explore use of progressively decreasing weights instead of a simple average. We find that the sentiment calculated using any of these measures is highly correlated with a simple average of Top 25 comments ($\rho \geq 0.88$).

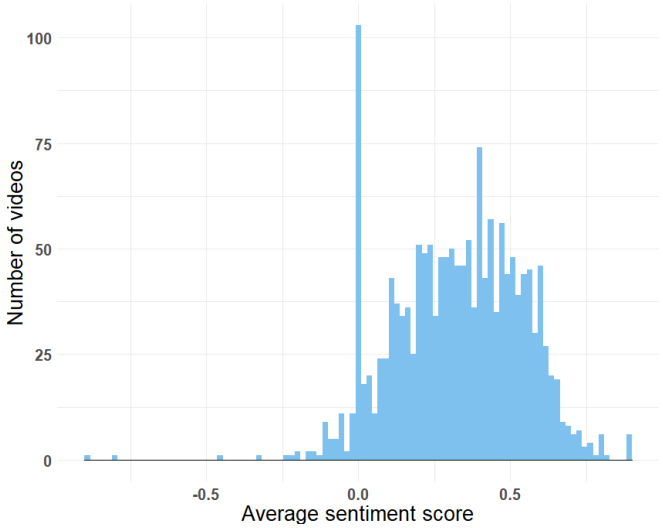

Figure 3 – Distribution of average sentiment score across Top 25 comments

Hence, we have a total of four continuous outcomes and one binary outcome. We find that the Pearson correlation coefficient between all outcomes ranges from 0.02 to 0.66 with a median of 0.20 (absolute value) as shown in Table 2, indicating that each measure potentially captures different underlying constructs.

	Log views	Log engagement	Log popularity	Log likeability	Binary Sentiment
Log views	1				
Log engagement	0.04	1			
Log popularity	0.20	0.66	1		
Log likeability	0.43	0.14	0.57	1	
Sentiment (binary)	−0.21	−0.15	0.02	0.15	1

Table 2: Correlation between outcomes

3.4 Features

Next, we generate features from the data scraped in Table 1 and list them in Table 3. As can be seen from the table, we have 33 fixed effects for channel, 11 fixed effects for category, features for video length, tags and playlist information, six time-based-covariates and an indicator variable for whether captions are available for the video. For the video description, a maximum of 160 characters are visible in Google Search and even fewer characters are visible below a YouTube video before the 'Show More' link (Cournoyer, 2014). Hence, we truncate each description to the first 160 characters as it is more likely to contribute to any potential association with our outcome variables. Captions are only present in 74% of

videos, and for those videos without a caption, we use Google's Cloud Speech-to-Text Video Transcribing API to transcribe the first 30 seconds of the audio file to English.[16]

Type	Class	Features
Structured Features	Fixed Effects	Influencer Fixed Effects (33)
	Fixed Effects	Category Fixed Effects (11)
	Length	Video Length (min)
	Tags	Number of video tags
	Playlist Information	Number of playlists the video is a part of
		Average position in playlist
		Average number of videos on all the playlists the video is a part of
	Time based covariates	Time between upload: Upload time and scrape time
		Year of upload (2006 to 2019)
		Time between upload: Given video and preceding video in master list
		Time between upload: Given video and succeeding video in master list
		Rank of video in master list
		Day fixed effects in EST (7) and Time of day fixed effects in intervals of 4 hours from 00:00 hours EST (6)
	Captions Indicator	Indicator of whether video has closed captions
Unstructured features	Text	Title
		Description (first 160 characters)
		Captions or Transcript (first 30 sec)
	Audio	Audio file (first 30 sec)
	Images	Thumbnail
		Image frame at 0 sec (first frame), 7.5 sec, 15 sec, 22.5 sec, 30 sec
Structured features	Complete Description	Total number of URLs in description
		Indicator of Hashtag in description

Table 3: Video features - structured and unstructured

We begin by focusing on the first 30 seconds for two reasons.[17] First, the minimum duration of video content that needs to be viewed for an impression to be registered is 30 seconds, and second, higher computational costs associated with more data in our deep learning models require us to restrict data size to a feasible amount. Similarly, we restrict the duration of the audio file to the first 30 seconds. We use audio data in addition to captions/transcript to analyze the presence of other sound elements such as music

[16] While most videos have English speech, if the first 30 seconds of audio have only non-English speech or there is only background music/sound, the transcription process results in an empty file. 65% of the 26% of videos that are transcribed result in an empty file.

[17] Note that as part of our robustness checks, we contrast our approach with the use of data from the middle 30 sec and last 30 sec of the video (see Section 5.4).

and animal sounds. Image data comprise high-resolution images that are 270 pixels high and 480 pixels wide. These images comprise the thumbnail and video frames at 0 sec (first frame), 7.5 sec, 15 sec, 22.5 sec and 30 sec.[18] We restrict our analysis to the first 30 seconds of the video to be consistent with our analysis of text and audio data, and we consider a maximum of five frames in the first 30 seconds because of computational constraints of GPU memory that can be achieved at a low cost.

We create two additional structured features from the complete description to use for analysis as the complete description is not supplied to the Text model. These features are included as they can lead the viewer away from the video. They comprise total number of URLs in description and an indicator for hashtag in description. The first three hashtags used in the description appear above the title of the video (if there are no hashtags in the title), and clicking on it can lead the viewer away from the video to another page that shows similar videos (Google, 2020e).[19]

3.5 Brand Usage

We compile a list of popular global brands and a comprehensive list of brands with offices in USA. Three lists of Top 100 Global brands in 2019 are obtained from BrandZ, Fortune100 and Interbrand. To this, we add a list of more than 32,000 brands (with US offices) from the Winmo database. This is further combined with brand names identified by applying Google's Vision API - Brand Logo Detection on thumbnails and video frames (0s, 7.5s, 15s, 22.5s & 30s) in our sample of 1620 videos. From this combined list, we remove more than 800 generic brand names such as 'slices,' 'basic,' 'promise,' etc. that are likely to be used in non-brand related contexts. Using *regular expressions*, we identify a list of 250 unique brands that are used in different text elements of a video: video title, video description (first 160 characters) and video captions/transcript (first 30 sec). The Logo detection API provides a list of 51 unique brands that are used in image elements of the video – thumbnails and video frames. The percentage of videos that have a brand used in each video element is as follows: title – 11.2%, description (first 160 characters) – 36.8%, captions/transcript (first 30 sec) – 17.2%, thumbnails – 1.1% and video frames (across five frames in first 30sec) – 2.6%.[20] The distribution of the number of brand mentions in each text element is shown in Figure 4a, and the number of brand logos in each image element is shown in Figure 4b.

[18] As each video can be recorded at a different frame rate or with variable framing rates, we capture the frame equal to or exactly after the specified time point. For example, a video recorded at a fixed rate of 15 frames/sec will have a frame at 7.46 sec and 7.53 sec but not 7.50 sec - so we record the frame at 7.53 sec in place of 7.50 sec.

[19] We do not have information on how often a video was recommended to viewers by the YouTube recommendation algorithm. We discuss the potential impact of not observing this feature in Section 5.2.4.

[20] We do not study brand usage in the Top 25 comments below a video as an *outcome variable* because only about 5% of the comments across all 1620 videos have a brand mentioned.

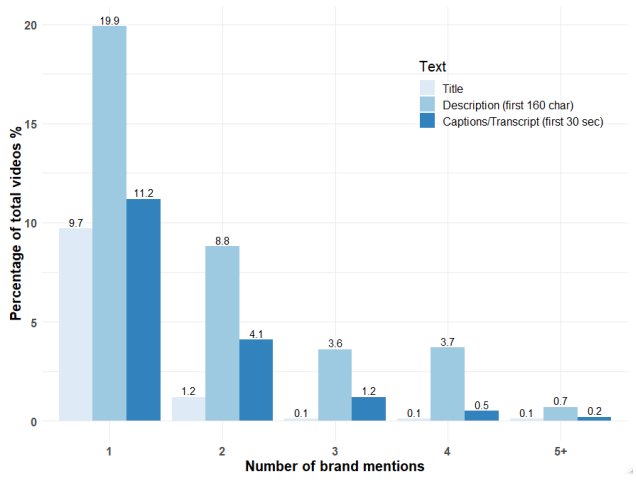

Figure 4a: Distribution of number of brand mentions in text

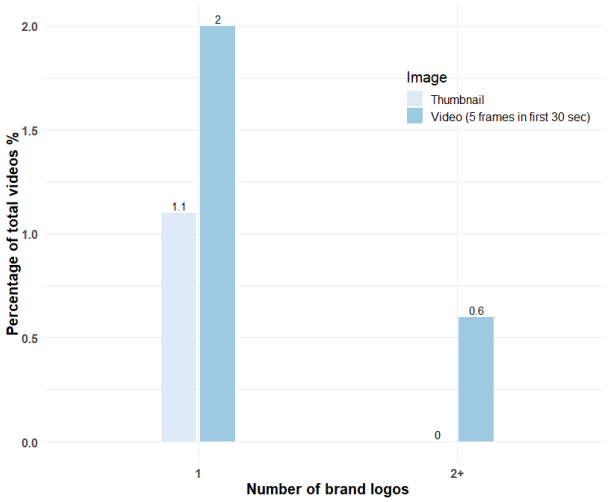

Figure 4b: Distribution of number of brand logos in images

We find that brand mentions are most common in the description (first 160 characters), followed by captions/transcript (first 30 sec), and then video title. Moreover, all text elements typically have only one brand mentioned once; the observations where two or more brands are mentioned include cases of the same or a different brand being mentioned again. Similarly, thumbnails and video frames (five equally spaced frames in the first 30 sec) typically have only one brand logo, but they comprise a very small percentage of the total videos in our sample. Overall, we find that our sample of influencers allows us to capture sufficient advertising information in textual data.

Furthermore, the US Federal Trade Commission (FTC) has three main guidelines for influencers. First, influencers need to disclose information about brand sponsorship in the video itself and not just in the description of the video. Second, they are advised to use words such as "ad," "advertisement," "sponsored" or "thanks to 'Acme' brand for the free product" to indicate a brand partnership. Third, it is recommended that they disclose brand partnerships at the beginning than at the end of a video (FTC, 2020). Hence, we check for presence of the words "ad/s," "advertisement/s," "sponsor/s" or "sponsored" in the captions/transcript (first 30 sec).[21] We find that less than 1% of videos make such a disclosure in the first 30 seconds.[22]

[21] We do not check for the presence of words such as "free" because they are often used in other contexts such as "feel free," "gluten free," etc.

[22] YouTube also has guidelines for influencers. It requires all channel owners to check a box in video settings that says 'video contains paid promotion' if their video is sponsored (Google, 2020d). If this box is checked, a tag – "Includes Paid Promotion" is overlaid on a corner of the video for the first few seconds when the video is played on YouTube. While information about the presence of this "tag" cannot be scraped or downloaded with the video to the best of our knowledge, manually checking different videos on YouTube in our sample reveals that there is little compliance to this requirement.

4. Model

Deep learning models are especially suited to analyze unstructured data (text, audio and images) as they can efficiently capture complex non-linear relationships and perform well in prediction tasks (Dzyabura & Yoganarasimhan, 2018). Figure 5a shows the traditional deep learning approach that uses unstructured data (e.g., images from videos) to predict an outcome variable. Features self-generated by deep learning models are known to have better predictive ability than ex-ante hand-crafted features passed to deep learning models (Dzyabura et al., 2018; L. Liu et al., 2018; X. Liu et al., 2019). This is because ex-ante feature engineering is unable to neither identify a comprehensive set of important features nor capture all the underlying latent constructs. However, hand-crafted features allow interpretability of the captured relationships which is not possible with self-generated features created by traditional deep learning models.

In Figure 5b, we show our "interpretable deep learning" approach that avoids ex-ante feature engineering and instead uses ex-post interpretation to allow for both good predictive ability of outcomes and interpretation of the captured relationships. To prevent the model from overfitting when analyzing a moderate sized dataset, we use transfer learning approaches where a model that is pre-trained on a separate task with large amounts of data (at a high cost) can be fine-tuned for our different but related task. This not only helps prevent overfitting but also aids in interpretation of the captured relationships. We use state-of-the-art model architectures with novel customizations that allow visualization of the captured relationships. Next, we describe each of the deep (transfer) learning models in more detail.

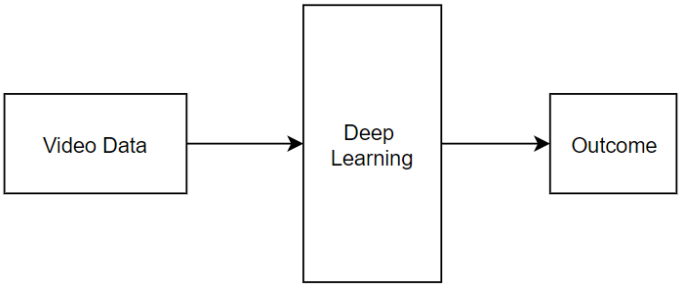

Figure 5a: Traditional Deep Learning Approach

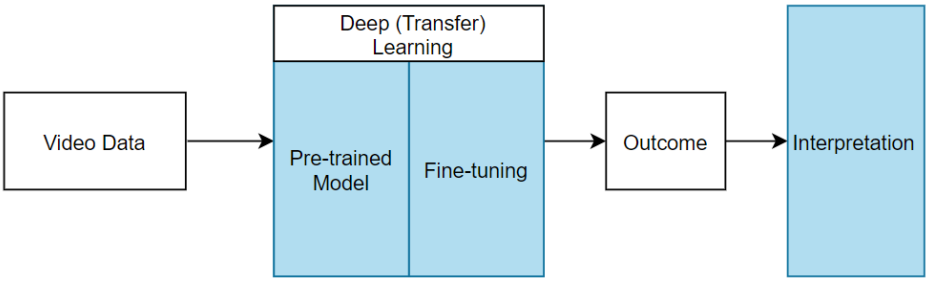

Figure 5b: Interpretable Deep Learning Approach

4.1 Text Model

Text data are analyzed using Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2018), a state-of-the-art NLP model that borrows the Encoder representation from the Transformer framework (Vaswani et al., 2017). The model is pre-trained using Book Corpus data (800M words) and English Wikipedia (2,500M words) to predict masked words in text and the next sentence following a sentence. Devlin et al. (2018) complete the pre-training procedure in four days using four cloud Tensor Processing Units (TPUs). The BERT model is fine-tuned to capture the association with our five outcomes using the framework shown in Figure 6.

The model converts a sentence into word-piece tokens[23] as done by state-of-the-art machine translation models (Wu et al., 2016). Furthermore, the beginning of each sentence is appended by the 'CLS' (classification) token and the end of each sentence is appended by the 'SEP' (separation token). For example, the sentence 'Good Morning! I am a YouTuber.' will be converted into the tokens ['[CLS]', 'good', 'morning', '!', 'i', 'am', 'a', 'youtube', '##r', '.', '[SEP]']. A 768-dimensional initial embedding learnt for each token during the pre-training phase is passed as input to the model, and is represented by the vector x_m in Figure 6, where m is the number of tokens in the longest sentence[24]. The token embedding is combined with a positional encoder t_m that codes the position of the token in the sentence using sine and cosine functions (see Devlin et al. (2018) for details). This is passed through a set of 12 encoders arranged sequentially. The output of the 'CLS' token is passed through a feed forward layer that is initialized with pre-trained weights from the next sentence prediction task, and has a tanh activation function, i.e. $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. This is followed by the output layer, which has a linear activation function, i.e. $linear(x) = x$, for the four continuous outcomes and a sigmoid activation function, i.e. $sigmoid(x) = \frac{e^x}{1+e^x}$, for the binary outcome sentiment.

[23] We use the BERT-base-uncased model (that converts all words to lower case and removes accent markers) as compared to the cased model, because the uncased model is known to typically perform better unless the goal is to study case specific contexts such as 'named entity recognition' and 'part-of-speech tagging'.

[24] Rare characters including emojis are assigned an 'UNK' (unknown) token and sentences shorter than the longest sentence are padded to the maximum length by a common vector.

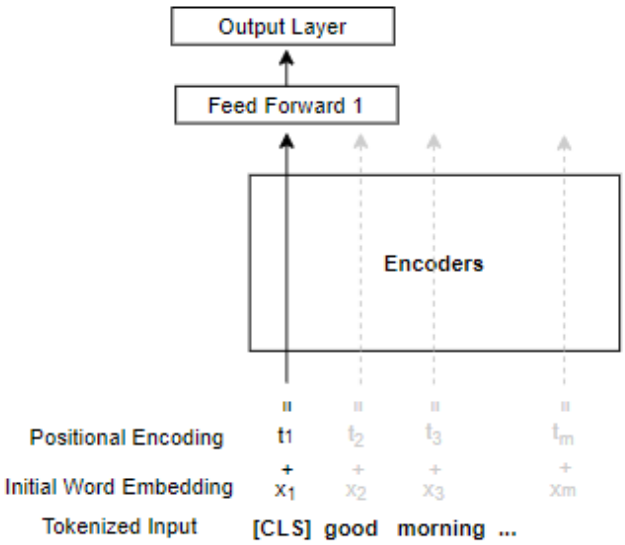

Figure 6: BERT Model Framework

In Appendix A, we explain the architecture of the Encoders which also contain the self-attention heads. The self-attention heads help the model capture the relative importance between word-pieces while forming an association with the outcome of interest. The three main advantages of BERT over conventional deep learning frameworks such as (Bidirectional) LSTM, CNN and CNN-LSTM that use word embeddings such as Glove and word2vec are as follows: (1) BERT learns contextual token embeddings (e.g., the embedding for the word 'bark' will change based on the context in which it used, such that the model can understand whether the word is referring to a dog's bark or a tree's outer layer) (2) The entire BERT model with hierarchical representations is pre-trained on masked words and a next sentence prediction task, thus making it suitable for transfer learning; whereas the conventional models only initialize the first layer with a word embedding (3) BERT uses a self-attention mechanism that allows the model to simultaneously (non-directionally) focus on all words in a text instead of using a sequential process that can lead to loss of information. These advantages are reflected in model performance when we compare it with conventional models (combinations of CNN and LSTM) in Section 5.1.

4.2 Audio Model

Audio data are analyzed using the state-of-the-art YAMNet model followed by a Bidirectional LSTM (Bi-LSTM) model with an attention mechanism, as shown in Figure 7. YAMNet takes the Mel-frequency spectrogram of the audio signal as input and passes it through a MobileNet v1 (version 1) model that is pre-trained on the AudioSet data released by Google (Gemmeke et al., 2017; Pilakal & Ellis, 2020).

YAMNet predicts sound labels from 521 audio classes[25] such as speech, music, animal, etc. corresponding to each 960ms segment of the audio file. The features from the last layer of the model, corresponding to the predicted audio classes, are passed through a Bi-LSTM layer with an attention mechanism to capture the sequential structure of sound. The model is then fine-tuned to capture associations with our five outcomes.

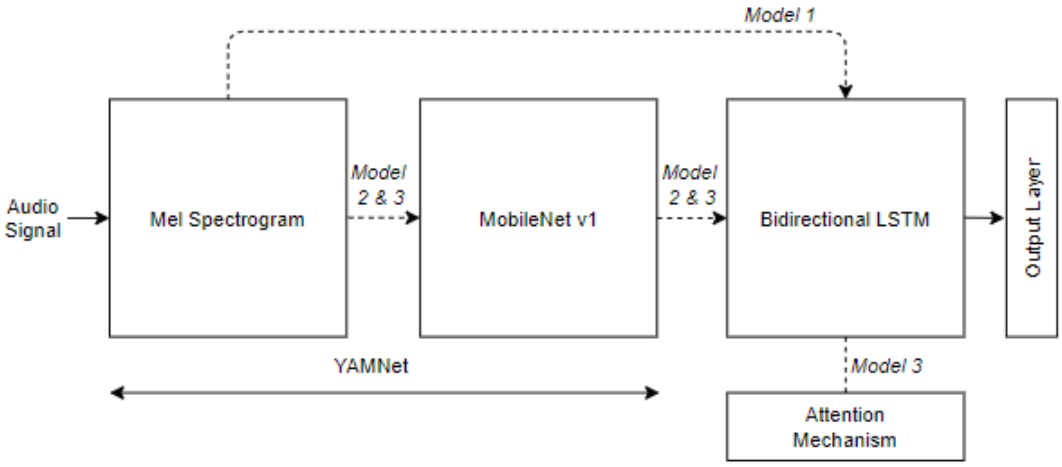

Figure 7: Audio Model

Next, we explain the model framework in more detail. As mentioned earlier in Section 2, we analyze the first 30 seconds of audio in each video file. Each 30 second audio clip is resampled at 16,000 Hz and mono sound, which results in 480,000 data points for each clip. To summarize the large number of data points, we generate a spectrogram that spans the frequency range of 125 to 7500Hz (note that the 2000-5000 Hz range is most sensitive to human hearing (Widex, 2016)) over which the YAMNet model has been pre-trained. This frequency range is then divided into 64 equally spaced Mel bins on the log scale, such that the sounds of equal distance on the scale also sound equally spaced to the human ear.[26] Each segment of 960ms from the spectrogram output, i.e., 96 frames of 10ms each with overlapping patches (that have a hop size of 490ms) to avoid losing information at the edges of each patch is passed as input to the MobileNet v1 architecture. The MobileNet v1 (explained in more detail in Appendix B) processes the spectrogram through multiple mobile convolutions which results in 521 audio class predictions across 60 moments (time steps) in the clip. The <521x60 > dimensional vector is then passed as input to the Bi-Directional LSTM layer with an attention mechanism (explained in more detail in Appendix B). This layer is made Bidirectional to allow it to capture the interdependence between

[25] The AudioSet data has more than 2 million human-labelled 10 sec YouTube video soundtracks (Gemmeke et al., 2017). Pilakal and Ellis (2020) remove 6 audio classes (viz. gendered versions of *speech* and *singing*; *battle cry*; and *funny music*) from the original set of 527 audio classes to avoid potentially offensive mislabeling. YAMNet has a mean average precision of 0.306.

[26] The spectrogram uses the pre-trained Short-Term Fourier Transform window length of 25ms with a hop size of 10ms that results in a 2998 x 64 (time steps x frequency) vector corresponding to 30 seconds of each audio clip.

sequential audio segments from both directions. For example, the interdependence between the sound of a musical instrument at 5 seconds and the beginning of human speech at 15 seconds can be captured by the model bidirectionally. We adopt the attention mechanism used for neural machine translation by Bahdanau et al. (2014) to help the Bi-LSTM model capture the relative importance between sound moments in order to form an association with an outcome of interest. The output of the Bi-LSTM with attention mechanism is passed through an output layer which has linear activations for the continuous outcome and sigmoid activation for the binary outcome. We compare the performance of this model framework (Model 3) with a model devoid of the attention mechanism (Model 2) and a model devoid of both the attention mechanism and MobileNet v1 (Model 1), in Section 5.1.

4.3 Image Model

Individual images are analyzed using the state-of-the-art image model – EfficientNet-B7 (Tan & Le, 2019) that has been pre-trained with "noisy student weights"[27] (Xie et al., 2019). This model not only has a high Top-5 accuracy on ImageNet (98.1%) but is also known to better capture salient regions of images as it uses compound scaling. It is also a relatively efficient model that uses only 66M parameters (and hence the name EfficientNet) as compared to other high performing models that use 8x times the number of parameters (Atlas ML, 2020). All our images (frames) are at a high resolution of 270 by 480 pixels which is the largest common resolution size available across all thumbnails and video frames in the dataset. Thumbnail images are passed as input to one EfficientNet-B7, and its final layers are fine-tuned to capture relationships with an outcome. The architecture of the EfficientNet-B7, whose main building block is the Mobile Inverted Bottleneck Convolution, is explained in detail in Appendix C. We compare the performance of the (pre-trained) EfficientNet-B7 with a 4-layer CNN model in Section 5.1.

As mentioned in Section 3, we analyze a maximum of five video frames in the first 30 seconds of each video, i.e., frames at 0s (first frame), 7.5s, 15s, 22.5s and 30s. Each image frame $i = 1$ to m, where m has a maximum value of 5, is passed through an EfficientNet-B7 model, and then the outputs from all the models are combined before passing it through an output layer. This is illustrated using the diagram in Figure 8.

[27] Xie et al. (2019) learn these weights by first pre-training the model on more than 1.2M labelled images from the ImageNet dataset (Russakovsky et al., 2015), and then use this trained model as a teacher to predict labels for a student model with 300M unlabeled images from the JFT Dataset (Xie et al., 2019). The two models are then combined to train a larger student model which is injected with noise (e.g., dropout, stochastic depth and data augmentation), and is then used as a teacher to predict labels for the original student model. This process is then iterated a few times to produce the EfficientNet-B7 model with pre-trained weights.

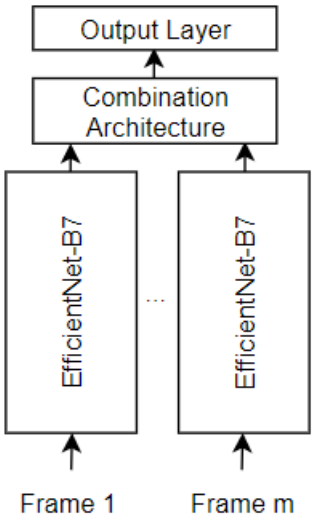

Figure 8: Image Model (Video Frames)

We compare the performance of four different 'combination architectures' that combine the outputs from each EfficientNet-B7. Two of the architectures are the best performing ones in Yue-Hei Ng et al. (2015), namely *Bi-LSTM*[28] and Max Pooling followed by Global Average Pooling (*Max-GAP*). The remaining two architectures are variants not tested by Yue-Hei Ng et al. (2015), namely Global Average Pooling followed by Max Pooling (*GAP-Max*) and Concatenation of Global Average Pooling (*C-GAP*). The Bi-LSTM architecture captures sequential information across the video frames, while the remaining three architectures preserve the spatial information across the video frames. The output of the combination architecture is passed through an output layer which has linear activations for the continuous outcome and softmax activation for the binary outcome. We explain the combination architectures in more detail in Appendix C.

4.4 Combined Model

We use the framework shown in Figure 9 to combine information from each of the unstructured models with the structured features, X_{it}, listed earlier in Table 3. The predicted outcome values, $\hat{Y}_{it}$ for video t by influencer i, from the best performing model for each unstructured feature, are fed into the combined model in addition to the structured features, X_{it}. This can also be represented by the following equation:

$$Y_{it} = g\left(X_{it}, \hat{Y}_{it\,Title}, \hat{Y}_{it\,Description}, \hat{Y}_{it\,Caption/Transcript}, \hat{Y}_{it\,Audio}, \hat{Y}_{it\,Thumbnail}, \hat{Y}_{it\,Video\ Frames}\right) + \epsilon_{it} \quad (1)$$

where Y_{it} is the observed outcome for video t by influencer i, g is the combined model used and ϵ_{it} is the error term. We test the performance of seven different combined models in Section 5.1. The combined

[28] While Yue-Hei Ng et al. (2015) use the LSTM approach, we use the Bidirectional LSTM (Bi-LSTM) as it can only perform better than LSTM.

models comprise four commonly used linear models – OLS[29], Ridge Regression, LASSO and Elastic Net, and three non-linear models – Deep Neural Net, Random Forests and Extreme Gradient Boosting (XGBoost) – that are known to capture non-linear relationships well.

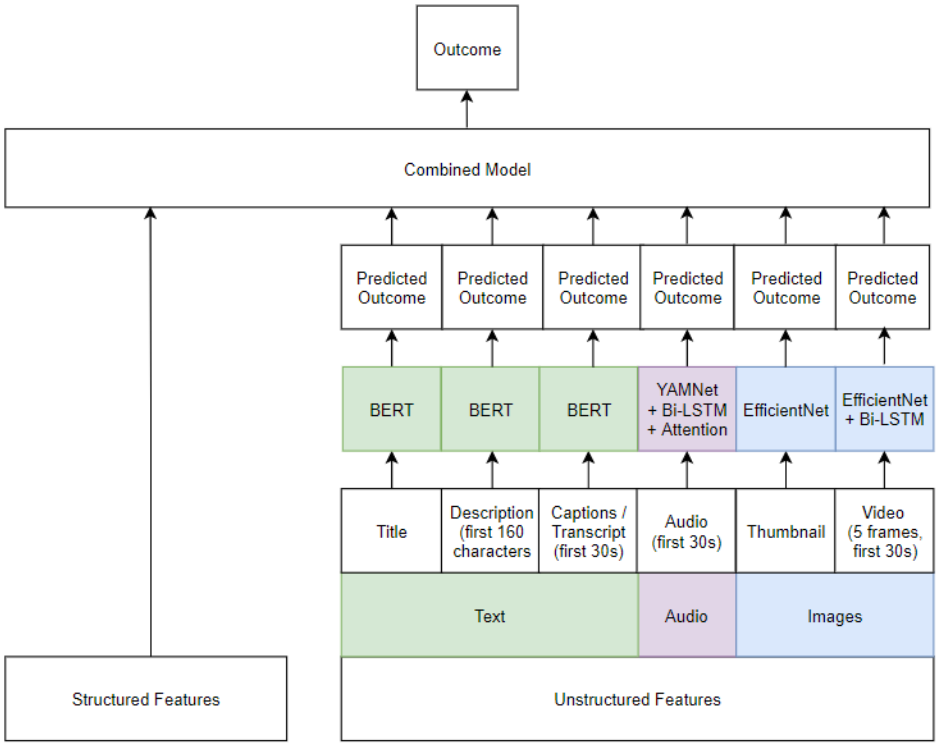

Figure 9: Combined Model

5. Results

In this section, we first detail the results on prediction and then on interpretation. We then dig deeper to see if we find patterns consistent with influencers "learning" about what makes their videos more engaging. We also carry out a robustness check where we estimate our model on video slices from the middle and end of videos as opposed to the beginning.

5.1 Prediction Results

We divide our random sample of 1620 videos into a 60% training sample (972 videos), 20% validation sample (324 videos) and 20% holdout sample (324 videos). We train the model on the training sample, tune the number of steps of Adam gradient descent on the validation sample, and then compute model

[29] We drop the multicollinear category fixed effects in OLS. We retain these fixed effects in the other models so that we can capture their relative importance with influencer fixed effects.

performance on the holdout sample.[30] First, we compare the predictive performance of each model with benchmarks models for the continuous outcome (views) and binary outcome (sentiment), and then apply the best performing model on the other three continuous outcomes (interaction rates).[31] In Appendix D, we compare our models with various benchmarks used in marketing literature. The Text model (BERT) performs better than benchmarks such as LSTM, CNN (X. Liu et al., 2019), CNN-LSTM (Chakraborty et al., 2019) and CNN-Bi-LSTM. The Audio model (YAMNet+Bi-LSTM+Attention) performs better than benchmark models devoid of the attention mechanism, thus demonstrating the benefit of capturing relative attention weights. The Image model (EfficientNet-B7) performs better than a conventional 4-layer CNN. Furthermore, the Bi-LSTM architecture which captures the sequential information from video frames performs better than other models that capture only spatial information.[32] Overall, we demonstrate that our models perform better than benchmark predictive models in the marketing literature and thus we do not compromise on predictive ability.

Table 4 summarizes the results from the best performing models for each component of unstructured data, which are now applied to all five outcomes. The Text model (BERT) predicts all the continuous outcomes with a low RMSE, (e.g., Title can be used to predict views within an average RMSE range of $\pm e^{1.66} = \pm 5.26$ views) and the binary outcome with moderately high accuracy (e.g., Title can predict sentiment with an accuracy of 72%). The Audio model can make predictions at a slightly poorer level of performance than the Text model. The Thumbnail model is unable to predict the continuous outcomes as well as the Text and Audio Model but performs better than the Audio model in predicting sentiment. The Video Frames model performs better than the Thumbnail model but poorer than Audio and Text models in predicting the continuous outcomes but performs comparably with the Thumbnail model in predicting sentiment. Overall, the prediction results show low RMSE that ranges from 0.71 to 3.09 when predicting the (log transformed) continuous outcomes (views, engagement, popularity and likeability) and moderately high accuracies ranging from 65% to 72% when predicting sentiment.

[30] We carry out our analysis using one NVIDIA Tesla P100 GPU (with 16GB RAM) from Google.

[31] We do not use a Multi-Task Learning (MTL) approach to simultaneously predict all five outcomes for two reasons. First, our final goal is to interpret the relationship between each individual outcome and video data (detailed in Section 5.2), which will not be possible with a MTL approach. Second, there is low to moderate correlation between all five outcomes as shown earlier in Table 2, which suggests that each outcome is capturing different underlying constructs.

[32] We also find that using five frames results in slightly improved performance than a model that uses only three or two frames.

Model	Data	Views	Sentiment	Engagement	Popularity	Likeability
BERT	Title	1.66	0.72	0.82	0.71	0.85
	Description (first 160c)	1.57	0.69	0.88	0.72	0.93
	Captions/transcript (first 30s)	1.75	0.70	0.92	0.76	0.99
YAMNet + Bi-LSTM + Attention	Audio (first 30s)	1.97	0.65	0.93	0.80	1.02
EfficientNet-B7	Thumbnail	3.09	0.68	1.75	1.34	1.43
EfficientNet-B7 + Bi-LSTM	Video Frames (0s,7.5s,15s,22.5s,30s)	2.23	0.68	0.97	0.80	1.03

Table 4: Best performing model for each component of unstructured data in holdout sample (RMSE for Views, Engagement, Popularity and Likeability; Accuracy for Sentiment)

The results of the Combined Model in Section 4.4 demonstrate that Ridge Regression has the best performance on the holdout sample for all the continuous outcomes (lowest RMSE) and also the binary outcome (highest accuracy) (see Appendix D for details). This suggests that structured features do not have substantial non-linear interactions with each other or with the predictions from the Text, Audio and Image models. Moreover, we find that the combined model of Ridge Regression has lower RMSE or higher accuracy than the results of the individual models in Table 4, suggesting that the holistic influence of all features is better than the individual influence of each unstructured feature. Next, we take the magnitude of each estimated coefficient from the Ridge Regression model applied on the training sample and scale it by the sum of the magnitude of all coefficient values which gives us the percentage contribution of each feature. We thus capture the relative importance or predictive power of a feature for each of the outcomes of interest while controlling for the presence of other features. The importance of each feature set is shown in Table 5.[33]

[33] Note that we scale all the features by their *L*2 norm before running the model so that we can make relative comparisons. Also, we sum up the coefficient values that lie within a class (e.g. sum up the coefficient values of influencer fixed effects, sum up the coefficient values of features of playlist information, etc.) to get an overall picture of the contribution of a class of features in predicting an outcome of interest.

Sr No.	Name	Views	Sentiment	Engagement	Popularity	Likeability
1	Influencer Fixed Effects	21.33%	18.74%	2.65%	12.71%	49.06%
2	Title	15.85%	15.07%	43.25%	22.97%	11.21%
3	Description (first 160c)	13.82%	14.39%	34.53%	17.88%	11.14%
4	Time based covariates	12.33%	7.98%	1.18%	11.20%	5.70%
5	Playlist Information	9.87%	0.32%	2.53%	5.72%	3.60%
6	Captions/transcript (first 30s)	9.77%	12.67%	7.49%	17.34%	2.91%
7	Total URLs in description	5.32%	0.06%	0.22%	3.84%	2.10%
8	Category Fixed Effects	4.92%	12.21%	0.48%	3.29%	10.42%
9	Thumbnail	2.53%	4.98%	2.04%	0.54%	1.53%
10	Video Length	2.15%	0.09%	0.38%	0.72%	0.61%
11	Audio (first 30s)	1.23%	5.21%	4.76%	2.76%	0.06%
12	Tags Count	0.68%	0.07%	0.18%	0.59%	0.21%
13	Captions Indicator	0.11%	2.34%	0.02%	0.04%	0.51%
14	Hashtag Indicator in Description	0.08%	0.14%	0.04%	0.29%	0.92%
15	Video Frames (0s,7.5s,15s,22.5s,30s)	0.001%	5.73%	0.27%	0.09%	0.02%

Table 5: Importance of features based on the Ridge Regression Model

We highlight the unstructured features in gray. Title, description (first 160 characters) and captions/transcript (first 30 sec) contribute relatively more than the other unstructured features in predicting all five outcomes. The ordering of relative influence, where we control for the presence of other structured and unstructured features, is the same as our finding in Table 4 where we do not control for the presence of other features. However, the results of the combined model especially allow us to make relative comparisons between outcomes. We find that title and description (first 160 characters) contribute most towards predicting engagement; captions/transcript (first 30s) contribute most towards predicting popularity, whereas thumbnail, audio (first 30s) and video frames (0s,7.5s,15s,22.5s,30s) contribute most towards predicting sentiment.

5.2 Interpretation Results

In this section, we interpret the results from each of the best performing transfer learned deep learning models. We interpret results using the predictions on the holdout sample, and not the training sample, so that the identified relationships are more likely to generalize out-of-sample. On the holdout sample, we focus on the following video elements in unstructured data. First, we focus on brand presence in text as it is of interest to brand sponsors and because past literature on influencer videos has not studied this (as mentioned earlier in Section 2.1). Second, we study audio elements such as duration of speech, music, and animal sounds. As past advertising literature has found that ads featuring animals and those using

voice-over and music reduce irritation towards the ad (Pelsmacker & Van den Bergh, 1999), we study the role of these audio elements in influencer videos. Similarly, we study image elements such as size of brand logos, clothes and accessories and persons as it is of interest to brand sponsors and has not been studied in past influencer marketing literature.

We divide our interpretation strategy into two steps and visually illustrate this in Figure 10. In Step 1 (Attention or Importance), we find relationships between predicted attention weights (gradients) and video elements that demonstrate a significant positive relationship (significant relationship). Doing so allows us to identify whether video elements play an important role in predicting an outcome. However, finding important elements here need not indicate that a change in the value of the element has a significant association with the outcome because of potential spurious relationships captured by the model (Vashishth et al., 2019). In Step 2 (Correlation), we identify significant relationships between video elements and the predicted outcomes of interest.[34] However, finding significant relationships here need not mean that the elements are important in order to predict the outcome because of confounds unassociated with attention paid to video elements. Hence, we find relationships that fall at the intersection of Step 1 and Step 2. Doing so allows us to identify relationships between video elements and outcomes that are also supported by a significant change in attention to video elements. Thus, we are able to generate a smaller set of hypotheses for formal causal testing. Next, we detail our interpretation strategy using the results from the Text, Audio and Image models.

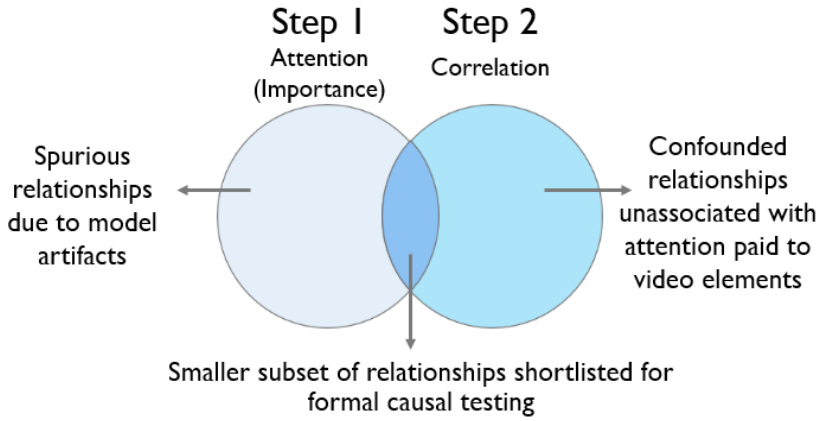

Figure 10: Interpretation strategy on holdout sample

[34] We use predicted outcomes in the holdout sample and not the observed outcome because the predicted outcomes have been influenced by the attention weights (gradients) and hence are comparable with the analysis in Step 1. Furthermore, the low RMSE values and moderately high accuracies (found in Section 5.1) do not preclude the use of predicted outcomes for analysis.

5.2.1 Interpretation: Text Model

We average the output across all the attention heads in the last encoder of the BERT model, which results in an attention vector of dimension <324, k, k> where 324 is the number of observations in the holdout sample, and <k,k> corresponds to k weights for k tokens, where k equals the maximum number of tokens for a covariate type – title, description (first 160 characters) or captions/transcript (first 30s). As mentioned in Section 4.1, the first token for each example is the 'CLS' or classification token. We are interested in the attention weights corresponding to this token because the output from this token goes to the output layer (as shown earlier in Figure 6). Thus, we get at an attention weight vector of dimension <324, k>, where each observation has k weights corresponding to the 'CLS' token. Note that the sum of the relative attention weights for each observation is one.

After finding the predicted attention weights in the holdout sample, we implement Step 1 where we run a regression of the predicted attention weights on brand presence to answer the following question:

1) Brand Attention: Do brand names receive more attention?

$$\log\left(AttentionWeight_{itj}\right) = \alpha_i + \gamma X_{it} + \beta_1\left(BIT_{itj}\right) + \beta_2(LOTX_{it}) + \beta_3\left(TP_{itj}\right) + \epsilon_{itj} \qquad (2)$$

where, $AttentionWeight_{itj}$ is the weight for token j (excluding 'CLS','SEP' and padding tokens) in video t made by influencer i, α_i is influencer fixed effect, X_{it} is the same vector of structured features used earlier in equation (1)[35], and ϵ_{itj} is the error term. BIT_{itj} is a 'Brand Indicator in Token' variable denoting whether token j used by influencer i in video t is a part of a brand name, $LOTX_{it}$ is the Length Of Text in video t made by influencer i, and TP_{itj} is the Token Position of token j used by influencer i in video t. In addition to studying main effects in equation (2), we also study interaction effects in Appendix E.

While equation (2) helps study the effect of brand presence on attention to the token, we also want to study the association between brand presence and the five outcomes of interest. Now, the predicted outcomes from the BERT model would have been influenced by the relative attention weights between words. Hence, in Step 2, we run a regression to answer the following question:

2) Brand Presence: Is brand presence associated with the predicted outcome?

$$PredictedOutcome_{it} = \alpha_i + \gamma X_{it} + \beta_1(BITX_{it}) + \beta_2(LOTX_{it}) + \epsilon_{it} \qquad (3)$$

[35] Note that we do not include category fixed effects in the linear regression to avoid multicollinearity with influencer fixed effects.

where, $BITX_{it}$ is a 'Brand Indicator in Text' variable denoting whether the text in video t by influencer i has a brand, and ϵ_{it} is the error term. We study interaction effects in Step 2 in Appendix E. The values of the coefficients of interest in each of the above equations are shown in Table 6.

Model for	Data Type	Step 1 - Eq(2)	Step 2 - Eq(3)
		BIT	BITX
Views	Title	27.60*	-5.26
	Desc	24.14*	61.21*
	Tran	7.25	7.72
Sentiment	Title	-25.35*	-6.27
	Desc	-16.75W	-46.04
	Tran	36.70*	-80.81*
Engagement	Title	32.10*	-6.36
	Desc	6.29	4.57
	Tran	626.23*	5.01
Popularity	Title	15.32*	-9.96
	Desc	18.88*	5.47
	Tran	156.72*	9.94
Likeability	Title	39.83*	-11.69
	Desc	82.65*	16.36W
	Tran	127.48*	5.78

Table 6: Results of the Text Regression Models
(– Significant ($p < 0.05$); W – Weakly Significant ($0.05 \leq p < 0.1$))*

The table reflects results corresponding to each type of unstructured text data – Title, Desc (first 160 characters of description) and Tran (first 30 sec of captions/transcript), and the model for each of the five outcomes – views, sentiment, engagement, popularity and likeability. The values in the table reflect a percent change in the non-log-transformed outcome (e.g., views and not log(views)) when a covariate is present.[36] Significant results are suffixed by * ($p < 0.05$) and weakly significant results ($0.05 \leq p < 0.1$) are suffixed by W.

We highlight the cells in gray that correspond to both (a) a positive and significant effect on attention weights and (b) a significant effect associated with the predicted outcome. Such a two-step comparison allows us to filter out significant relationships confounded by factors unrelated to brand attention. Doing so allows us to identify relationships that are more likely to have causal effects when tested in the field. We find two main effects that are significant in both steps. First, brand mention in description (first 160 characters) is associated with an increase in attention and an increase in predicted views. Second, brand mention in captions/transcript (first 30s) is associated with an increase in attention but negatively associated with predicted sentiment. However, we do not find any significant evidence to

[36] Note that we run a logistic regression for sentiment instead of a linear regression (in Section 5.2.1, 5.2.2 and 5.2.3).

show that the effect of brand mentions can vary based on length of text or its position in the text (see Appendix E).

Next, we illustrate an example of how text data in a video in the holdout sample can be visually interpreted. In Figure 11, we show the attentions weights on the captions/transcript (first 30s) from a video of a tech & business influencer. The words are tokenized into word-pieces in the figure as done by the model, and a darker background color indicates relatively higher attention weights. As can be seen in Figure 11, on average more attention is paid to the brand 'iphone' than other tokens in the text.[37] The model predicts a 'not positive' sentiment for this clip, and this matches the observed sentiment as well. These findings can help influencers design content and test it to obtain causal effects in a field setting. Similarly, brands can evaluate content using these findings to determine sentiment.

Predicted sentiment: Not Positive
Observed sentiment: Not Positive

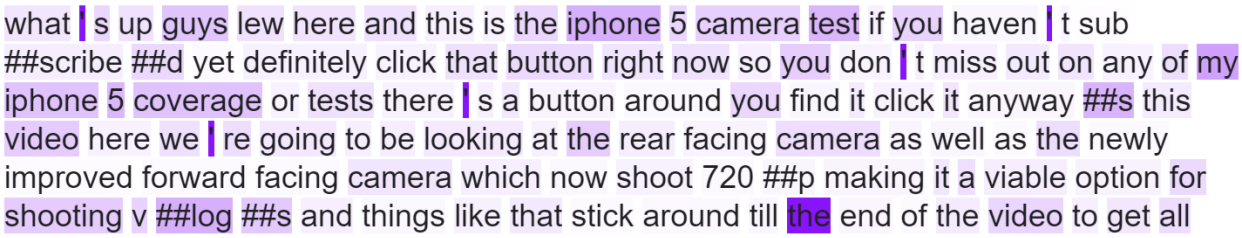

Figure 11: Attention Weights in captions/transcript (first 30s) of a video

5.2.2 Interpretation: Audio Model

The YAMNet model (Mel Spectrogram + MobileNet v1) finds the predicted probability of each moment of the 30 second audio clip belonging to 521 sound classes. A 30 second audio clip has 60 moments, where each moment is 960ms long, and the subsequent moment begins after a hop of 490ms. We divide the 521 sound classes into 8 categories based on the AudioSet ontology (Gemmeke et al., 2017) – Human (58.1%), Music (29.1%), Silence (3.5%), Things (0.7%), Animal (0.6%), Source Ambiguous (0.1%), Background (0.1%) and Natural (0.1%), where the percentage in brackets indicate the percentage of moments across our sample of 97,200 moments (1620 videos x 60 moments) that contain a sound of that category with probability greater than half. Note that 10.9% of moments are unclassified by the model; in addition, the same moment can be classified into multiple categories if sounds from two or more categories occur at the same moment (e.g., human speech while music is playing). The Audio Model – YAMNet + Bi-LSTM with attention, gives us 60 attention weights corresponding to each moment. Note

[37] Note that the model pays different attention to the word 'the' based on the context in which it is used.

that the sum of the relative attention weights for each observation is one. In Step 1, we run a regression of the predicted attention weights in the holdout sample to answer the following question:

1) Do certain moments of sound receive more attention?

$$\log\left(AttentionWeight_{itj}\right) = \alpha_i + \gamma X_{it} + \sum_{z=1}^{8} \beta_{1z}\left(CI(z)_{itj}\right) + \beta_2\left(CI(Human)_{itj} x\ CI(Music)_{itj}\right) + \beta_3\left(Location_{itj}\right) + \epsilon_{itj} \quad (4)$$

where, $AttentionWeight_{itj}$ is the weight for moment j in video t made by influencer i, α_i is influencer fixed effect, X_{it} is the same vector of structured features used earlier in equation (2), z = 1 to 8 corresponds to the 8 sound categories, $CI(z)$ is the Category Indicator for category z in moment j, and $CI(Human)$ x $CI(Music)$ corresponds to moments when both Human and Music sounds occur together, and $Location$ corresponds to location of the moment within the 60 moments of the audio clip, and ϵ_{itj} is the error term.

Next in Step 2, we examine whether these moments of sound have a significant effect on each outcome. We use the predicted outcomes from the Audio model as they would have been influenced by the relative attention weights between moments. We run a regression to answer the following question:

2) Are sound durations of certain sound categories associated with the predicted outcome?

$$PredictedOutcome_{it} = \alpha_i + \gamma X_{it} + \sum_{z=1}^{8} \beta_{1z}\left(Sum\ of\ CI(z)_{it}\right) + \beta_2(Sum\ of\ CI(Human)\ x\ CI(Music)_{it}) + \beta_3(Brand\ Indicator\ in\ Audio_{it}) + \epsilon_{it} \quad (5)$$

where, $Sum\ of\ CI(z)_{it}$ corresponds to the sum of the Category Indicator for z across the first 60 moments in video t made by influencer i, and $Sum\ of\ CI(Human)\ x\ CI(Music)_{it}$ finds the total duration when human and music sounds occur together, $Brand\ Indicator\ in\ Audio_{it}$ borrows the textual information in captions/transcript (first 30 sec) and acts as an indicator for whether a brand was mentioned in the audio clip, and ϵ_{it} is the error term. The results for the coefficients in each of the above equations are shown in Table 7. As three sound classes – Source Ambiguous, Background and Natural are present in only 0.1% of moments, we only use them as controls and hence their coefficients are not reported in the table. The values in the table reflect a percent change in the non-log-transformed outcome when a covariate is present (equation (4)) or increases by one unit (equation (5)).

	Model for	Category Indicator					Human x Music	Loc-ation	Brand Indic-ator
		Human	Music	Silence	Things	Animal			
Step 1 Eq(4)	Views	23.22*	18.48*	-17.61*	-31.36*	10.15W	-7.22*	-2.34*	NA
	Sentiment	-17.82*	36.47*	35.36*	0.97	21.82*	-19.07*	0.04*	NA
	Engagement	-0.54N	5.20*	-9.79*	6.58	-6.38	-15.29*	-1.65*	NA
	Popularity	-6.29*	36.66*	-22.03*	-2.74	1.89	- 4.98*	-1.35*	NA
	Likeability	6.62*	5.10*	-6.63*	-2.58	11.81*	0.33	0.26*	NA
Step 2 Eq(5)	Views	2.82*	-0.4	-1.66*	8.45*	0.30	-1.13*	NA	5.74
	Sentiment[38]	-1.43*	0.13*	0.08*	1.07*	0.14*	1.03	NA	-9.82*
	Engagement	-0.43W	-2.28*	-2.11*	-6.27*	-0.16	-0.13	NA	3.29
	Popularity	-0.62*	-1.66*	-1.24*	-6.33*	0.58	0.16	NA	2.84
	Likeability	0.17*	-0.26*	-0.02W	0.13W	0.25*	0.12*	NA	-0.19

Table 7: Results of the Audio Regression Models
(– Significant ($p < 0.05$); W – Weakly Significant ($0.05 \leq p < 0.1$))*

As done in Section 5.2.1, we highlight the cells in gray that correspond to both (a) a positive and significant effect on attention weights and (b) a significant effect on predicted outcome. This allows us to filter out significant relationships affected by confounds unassociated with an increase in attention to audio moments. We find nine significant results. First, human sounds (without simultaneous music) are associated with an increase in attention, and their longer durations are associated with higher predicted views and likeability. Second, music (without simultaneous human sounds) is associated with an increase in attention, and its longer duration is associated with lower predicted engagement, popularity and likeability but higher predicted sentiment. Last, animal sounds are associated with an increase in attention, and their longer durations are associated with higher predicted sentiment and likeability. In addition, we also find that brand presence (in first 30 seconds of audio) is associated with lower predicted sentiment (while controlling for duration of each class of sound), thus complementing our similar finding with the Text Model. Thus, we identify significant relationships between sounds and outcomes that are supported by significant increase in attention paid to audio moments.

Next, we illustrate an example of how attention paid to audio moments in a video in the holdout sample can be visually interpreted. We focus on the relationship between speech and music. In Figure 12, we show the first 30 seconds of the audio clip of a travel influencer using four sub plots. The first plot shows the variations in the amplitude of the 30 second audio wave (sampled at 16 KHz) followed by the spectrogram of the wave where brighter regions correspond to stronger (or louder) amplitudes. Next, we show the interim output of the Audio model with the top 10 sound classes at each moment in the audio, where the darker squares indicate higher probability of observing a sound of that class at that moment (Pilakal & Ellis, 2020). The last plot displays the attention weights corresponding to each moment in the audio clip, where the darker squares indicate higher relative attention placed on that moment while

[38] p values (confidence intervals) are calculated using penalized log likelihood instead of maximum log likelihood due to 'complete separation' in logistic regression.

forming an association with the outcome sentiment. As can be seen in the figure, relatively more attention is directed to moments where there is music but no simultaneous speech. The model predicts a positive sentiment for this clip, and this matches the observed sentiment as well.

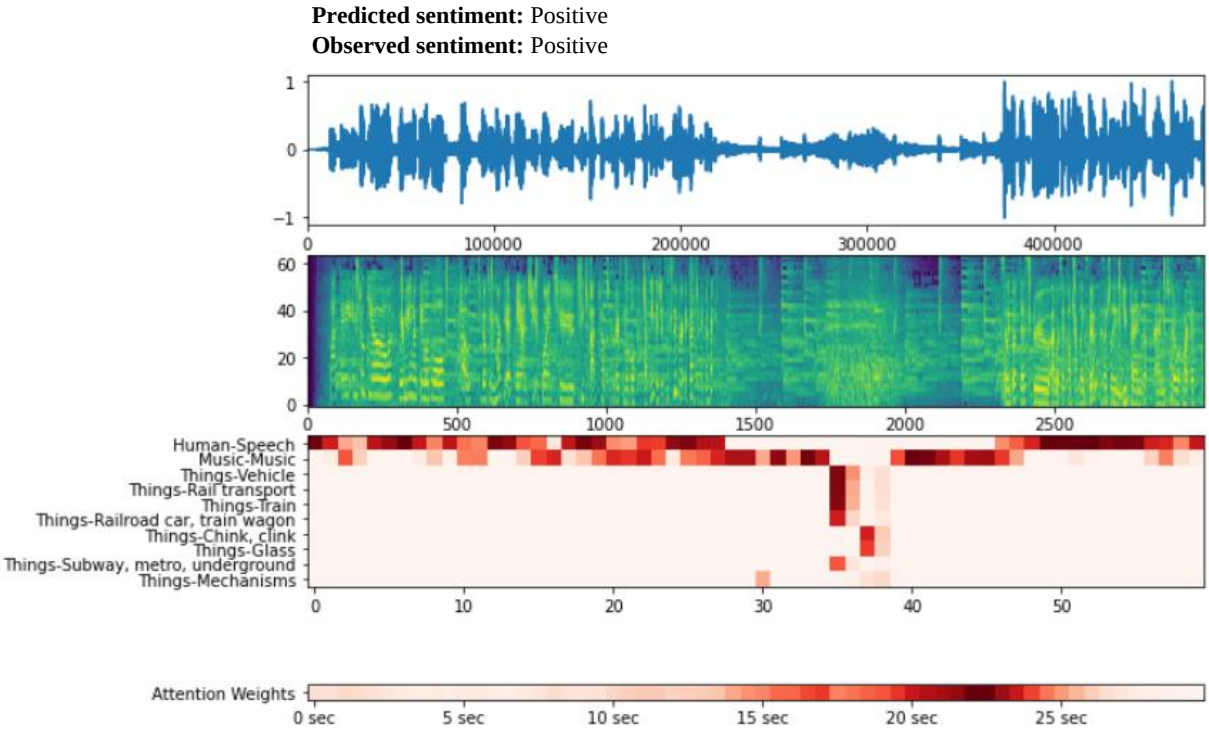

Figure 12: Attention weights in an audio clip (first 30s) of a video

5.2.3 Interpretation: Image Model

The salient parts of the images that are associated with an outcome of interest are visualized through gradient based activation maps (cf. Selvaraju et al., 2017). Gradients are found by taking the derivative between the continuous outcome (or class of predicted outcome for sentiment) and the output of the activation layer after the last convolution layer in the EfficientNet-B7 model. However, unlike Selvaraju et al. (2017), we do not apply the ReLU (Rectified Linear Unit) activation on the gradient values as we would like to retain negative gradient values for interpretation. In the Video Frame model, this process is carried out in each EfficientNet-B7 model corresponding to each video frame. Areas of the image with positive gradients correspond to regions that are positively associated with continuous outcomes and the predicted class of sentiment. To systematically identify and summarize the salient regions in thumbnails and video frames, we use Google's Cloud Vision API to detect objects and brand logos[39] in the images in

[39] We use a 70% confidence level of the Vision API to detect objects and a 90% confidence level to detect brand logos to be conservative in our estimates.

the holdout sample. The API returns the vertices of the identified item which allows us to create a rectangular bounding box to define its area. Next, we divide the identified objects into six categories – Persons (44.2%), Clothes & Accessories (30.9%), Home & Kitchen (11.0%), Animal (6.0%), Other Objects (3.7%) and Packaged Goods (1.9%); we let Brand Logos (2.3%) be the seventh category. The percentage in brackets indicates the percentage of items in that category out of a total of 4066 items (3973 objects + 93 brand logos) identified across 1944 frames in the holdout sample (324 videos x (1 thumbnail frame + 5 video frames)). In Step 1, we run a regression of the predicted gradient values in the holdout sample to answer the following question:

1) Is size of objects/brand logos associated with mean gradient over its area?

$$MeanGradientValues_{itz} = \alpha_i + \gamma X_{it} + \beta ItemSize_{itz} + \epsilon_{itz} \quad (6)$$

where, $MeanGradientValues_{itz}$ is the mean gradient values across the area (pixels) occupied by all items of category z in video t made by influencer i, α_i is influencer fixed effect, X_{it} is the same vector of structured features used earlier in equation (2), z = 1 to 7 corresponds to each item category, $ItemSize_{itz}$ is the percentage of full image size occupied by all items of category z in video t made by influencer i, and ϵ_{itz} is the error term.

Now, an increase in gradients is directly correlated with an increase in predicted values of continuous outcomes or the predicted class of binary outcome by design. However, we would like to eliminate spurious relationships due to model artifacts or confounds associated with the presence of other items in the image. Hence in Step 2, we run a regression to answer the question:

2) Is size of an object/brand logo associated with the predicted outcome?

$$PredictedOutcome_{it} = \alpha_i + \gamma X_{it} + \sum_{z=1}^{7} \beta_{1z}\, ItemSize(z)_{it} + \epsilon_{it} \quad (7)$$

where, $ItemSize(z)_{it}$ is the percentage of the full image size occupied by all items of category z in video t made by influencer i, and ϵ_{it} is the error term. We run this regression for thumbnails and for the average of five video frames (in the first 30 sec) for each of the five outcomes. The results for the coefficients in the above equations are shown in Table 8. The values in the table reflect a percent change in the non-log-transformed outcome when size of an item increases by one percent.

Equation	Model for	Data Type	Sub Covariate						
			Person	Clothes & Acc	Home & Kitchen	Animal	Other Objects	Packaged Goods	Brand Logos
Step 1 Eq(6)	Views	Thumbnail	0.03	0.06	0.18*	-0.00	0.05	0.02	1.01
		Avg 5 frames	0.57*	1.04*	0.85*	0.86*	0.60*	-0.93*	-2.56
	Sentiment	Thumbnail	-0.05*	-0.09*	-0.24*	-0.04	-0.07	0.01	-0.72
		Avg 5 frames	-0.03W	-0.07	-0.17*	0.09	-0.19W	-0.31	-4.6
	Engagement	Thumbnail	-0.08*	-0.06	-0.18W	-0.05	-0.02	-0.08	0.77
		Avg 5 frames	0.36*	0.75*	0.41*	0.89*	0.12	0.20	-1.45
	Popularity	Thumbnail	-0.11*	-0.19*	-0.14	-0.05	-0.11	-0.15	1.53*
		Avg 5 frames	0.51*	0.92*	0.71*	0.50*	0.41W	0.35	3.05
	Likeability	Thumbnail	-0.01	-0.02	0.06	-0.03	-0.02	0.05	1.19*
		Avg 5 frames	0.49*	0.80*	0.56*	0.61*	0.48*	-0.08*	1.60
Step 2 Eq(7)	Views	Thumbnail	1.06W	-0.74	2.56	1.55	0.97	-0.37	6.44
		Avg 5 frames	-0.00	0.01	-0.11*	0.00	-0.04	0.03	4.86W
	Sentiment	Thumbnail	-0.25	0.50	5.94	2.02	-7.00*	3.19	-100
		Avg 5 frames	-0.46	-4.66*	-2.72	23.97*	-6.73	5.95	40.27
	Engagement	Thumbnail	0.20	1.03*	-1.22	0.57	0.77	-1.34	-7.46
		Avg 5 frames	0.40*	0.66*	0.26	0.20	0.33	0.45	-1.54
	Popularity	Thumbnail	0.18	-0.21	1.02	0.32	0.84W	0.04	6.96
		Avg 5 frames	0.02	0.00	0.03	0.01	0.09W	0.07	0.64
	Likeability	Thumbnail	-0.01	-0.53	0.82	-0.65	-0.12	0.56	8.12
		Avg 5 frames	0.01	-0.03	0.02	-0.02	-0.03	0.15W	0.21

Table 8: Results of the Image Regression Models
(– Significant ($p < 0.05$); W – Weakly Significant ($0.05 \leq p < 0.1$))*

We highlight the cells in gray that correspond to a significant effect for both equations in the same direction. These cells show evidence of not only a significant effect on mean attention weights but also a significant effect in the same direction on predicted outcome while controlling for the presence of other items. We find that larger pictures of persons or clothes & accessories in video frames (first 30 sec) are associated with an increase in mean attention and an increase in predicted engagement. Influencers and brands promoting clothes & accessories are likely to benefit from testing this relationship for causal effects in a field setting.

Next, we illustrate how attention paid to image pixels on the video frames of a video in the holdout sample can be visually interpreted. We focus on the first three frames at 0 sec, 7.5 sec and 15 sec for a video of a gaming influencer in Figure 13. The first row shows the original frames which are overlaid with bounding boxes for items identified by the Vision API. The second row shows the heat map (positive gradient values) while forming an association with engagement. Brighter heat maps correspond to values that are more positively correlated with engagement. We find that pixels associated with images of persons have brighter heat maps (as compared to other parts of the image), and as the area occupied by the person decreases, the percentage of the area of the person that is salient also decreases. This conforms with the significant findings from Table 8. Furthermore, the predicted engagement for this example is 15

comments per 10,000 views which is less than the median engagement of 19 comments per 10,000 views, which can be expected given that the size of the person is progressively decreasing in subsequent frames.

Predicted Engagement: 15 comments per 10K views
Median Engagement: 19 comments per 10K views

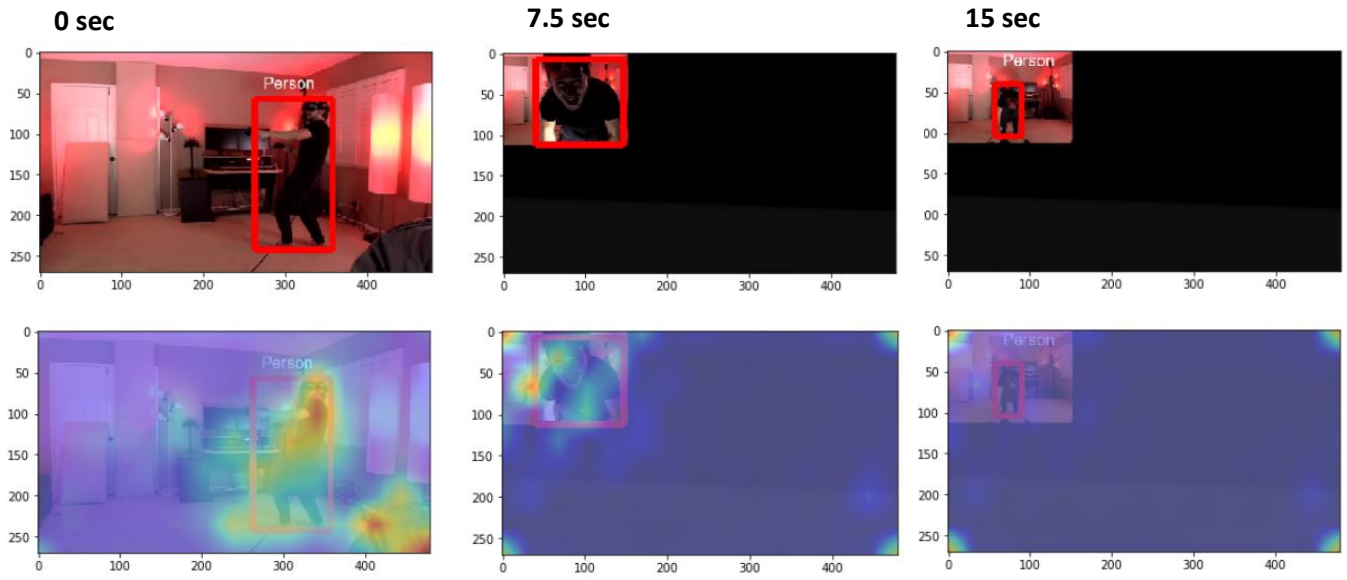

Figure 13: Gradient heat map (associated with engagement) in frames of a video

5.2.4 Summarizing Insights

We filter out 16 significant relationships affected by confounds unassociated with an increase in attention (change in gradients) to video elements (i.e., significant in Step 2 but not in Step 1). Next, we carry out a check to ensure that the results in Step 2 remain significant while controlling for the presence of other unstructured elements, using the following equation:

$$PredictedOutcome_{it} = \alpha_i + \gamma X_{it} + \beta_1(BITX_{it}) + \beta_2(LOTX_{it}) + \sum_{z=1}^{8} \beta_{3z}\,(Sum\ of\ CI(z)_{it}) + \beta_4(Sum\ of\ CI(Human)\ x\ CI(Music)_{it})) + \sum_{z=1}^{7} \beta_{5z}\ ItemSize(z)_{it} + \epsilon_{it} \quad (8)$$

Using the above equation, we eliminate one relationship that is no longer significant. Overall, we eliminate more than 50% of the relationships (out of 29 significant relationships in Step 2) and identify a smaller subset of 12 relationships (or hypotheses) that can be plausibly causal and tested in the field. We find the greatest number of significant results from the Audio model (8), followed by the Text model (2), and then the Image model (2) which are all summarized in Table 9. These significant associations between elements of one of the three modalities (text, audio or images) and marketing outcomes (views, interaction rates or sentiment) are likely not confounded by the presence of other modalities as we control for them. The effect sizes in Table 9 reflect a percent change in the non-log-transformed outcome when a

covariate is present (brand mentions in Step 1 & 2, audio moments in Step 1), increases by one unit (audio moments in Step 2) or increases by one percent (image sizes in Step 1 & 2). The effect sizes corresponding to the outcome in Table 9 reflect the coefficient sizes from equation (8).

	Significant *increase* in attention (A) and significant *increase* in outcome (O)			Significant *increase* in attention (A) but significant *decrease* in outcome (O)	
Outcome	Text Model	Audio Model	Image Model	Text Model	Audio Model
Views	brand mentions in description (first 160 char) A: 25.14% O: 64.63%	more speech (without simultaneous music) in first 30 sec of audio A: 23.22% O: 2.75%			
Sentiment		more music (without simultaneous speech) A: 36.47% O: 0.09% or more silence A: 35.36% O: 0.08% in first 30 sec of audio		brand mentions in first 30 sec of captions/transcript A: 36.70% O: - 89.10%	
Engagement			Larger pictures of persons A: 0.36% O: 0.38% or clothes & accessories A: 0.75% O: 0.62% in first 30 sec of video frames		more music (without simultaneous speech) in first 30 sec of audio A: 5.20% O: - 2.28%
Popularity					more music (without simultaneous speech) in first 30 sec of audio A: 36.66% O: - 1.67%
Likeability		more speech (without simultaneous music) A: 6.62% O: 0.17% or more animal sounds A: 11.82% O: 0.24% in first 30 sec of audio			more music (without simultaneous speech) A:5.10% O: - 0.26% in first 30 sec of audio

Table 9: Results from interpreting Regression Models

Finally, we highlight how the unobserved YouTube recommendation algorithm could potentially impact our findings. The algorithm analyzes watch history and video content to recommend videos with higher expected watch time for a viewer (Covington et al., 2016). In our analysis, if the ex-post elements that we study are correlated with unobserved features that cause higher expected watch time then our results corresponding to the outcome views need not be causal (because views can be expected to be highly correlated with watch time). However, the remaining four outcomes (sentiment, engagement, popularity and likeability) are unlikely to be highly correlated with watch time because of their low correlation with views (shown earlier in Table 2). Hence, the algorithm is unlikely to be a confounder for the significant relationships that we find for these four outcomes.

5.3 Learning Patterns

In this section we take a deeper dive to examine if influencer videos exhibit informal patterns suggesting that influencers are learning and acting on these relationships over time. To do this, we select three product categories that include influencers at both ends of the follower range ("micro" with < 100K subscribers and "mega" with ≥ 1M subscribers as per industry classification (Ismail, 2018)) and that have at least 100 videos in each group. These are Travel (2 micro and 1 mega), Parenting (1 micro and 1 mega) and Home (2 micro and 1 mega). We then choose a smaller sample of videos from each mega influencer corresponding to the total number of videos of the micro influencers in each category so that we have a balanced sample within each category. As before, we exclude those videos in which either likes, dislikes or comments were disabled by the influencer(s), leaving us with a total of 947 videos for Travel, 900 videos for Parenting and 322 videos for Home (all scraped in January 2020).

Our goal is to study whether influencers are changing their videos over time on the video elements that were found to have significant relationships with the outcomes in Table 9. Our identification strategy is a test of the change in the coefficient of "Video Number x Indicator of Influencer group" estimated separately for the first half and the second half of all videos uploaded by each influencer. These coefficients are obtained (for each of the three product categories) via the regression below.

$$
\begin{aligned}
V_{tih} = {} & \gamma_h Z_{tih} + \beta_{(h)1}\left(Indicator\ for\ Microinfluencer_{itp}\right) + \\
& \beta_{(h)2}\left(Indicator\ for\ Megainfluencer_{itp}\right) + \\
& \beta_{(h)3}\left(Video\ Number_{tip}\ x\ Indicator\ for\ Microinfluencer_{tip}\right) + \\
& \beta_{(h)4}\left(Video\ Number_{tip}\ x\ Indicator\ for\ Megainfluencer_{tip}\right) + \epsilon_{tip}
\end{aligned}
\tag{9}
$$

where V is the video element for video t by influencer i in half h, $h = \{1,2\}$. The video elements V were identified and listed in Table 9. Z includes controls for video length, number of tags, features from playlist, time between uploads, day and time of day fixed effects, captions indicator, number of URLs in description, and indicator of hashtag in description. $Video\ Number$ is the serial number of the video uploaded by the influencer, where a $Video\ Number$ of 0 corresponds to the first video uploaded by the influencer. To document whether there are patterns consistent with learning, we compare and contrast the coefficients $\beta_{(1)3}$ and $\beta_{(2)3}$ as well as $\beta_{(1)4}$ and $\beta_{(2)4}$ for micro and mega-influencers respectively. We find that only 3% of the relationships exhibit significant coefficient values for both $\beta_{(1)k}$ and $\beta_{(2)k}$, $k = \{3, 4\}$, suggesting that majority of micro and mega influencers across Travel, Parenting and Home categories do not exhibit patterns consistent with learning these relationships over time. While this analysis is fairly simple, we do not find any evidence of systematic changes. We leave a more detailed analysis of this topic for future research.

5.4 Analysis Using Other Slices of Influencer Videos

In this section, we analyze content in the middle 30 sec and last 30 sec of each video across transcript/captions, audio and images as a robustness check. Specifically, we compare our findings with the results for the first 30 sec of the video presented in Section 5.1. As shown in Table 10, using information from the middle or end of the video does not perform better than using information from the beginning for predicting all five outcomes on the holdout sample (using any of the three modalities - transcript/captions, audio or image frames). This suggests that the information in the beginning of the video is most important for predicting all outcomes. Furthermore, we also combine information from the beginning, middle and end of video for each modality of unstructured data using a Ridge Regression model (found as best performing in Section 5.1) to predict each of the five outcomes. We find that prediction results improve in only 5 out of the 15 cases (3 modalities x 5 outcomes) which demonstrates that information in the beginning of the video often captures variation in data explained by the middle and end of videos. Given that the predictions using these two sets of information do not dominate, and in the interest of parsimony and computational efficiency, we keep the focus of our analysis on the initial 30 sec of the videos.[40]

[40] We refer the interested reader to Appendix F for details on the interpretable results from the middle and end slices of videos.

Model	Data	Data Location	Views	Sentiment	Engagement	Popularity	Likeability
BERT	Captions/ transcript (30s)	Beginning	1.75	0.70	0.92	0.76	0.99
		Middle	1.92	0.67	0.98	0.80	1.03
		End	1.88	0.68	0.98	0.79	1.05
YAMNet + Bi-LSTM + Attention	Audio (30s)	Beginning	1.97	0.65	0.93	0.80	1.02
		Middle	2.26	0.62	0.96	0.80	1.02
		End	2.27	0.63	0.96	0.81	1.02
EfficientNet-B7 + Bi-LSTM	Video Frames (five equally spaced in 30s)	Beginning	2.23	0.68	0.97	0.80	1.03
		Middle	2.31	0.65	1.01	0.83	1.03
		End	2.31	0.68	0.99	0.82	1.06

Table 10: Predictive accuracy in holdout sample using unstructured data from beginning, middle and end of videos (RMSE for Views, Engagement, Popularity and Likeability; Accuracy for Sentiment)

6. Implications for Influencers and Marketers

In this section, we illustrate how our approach and findings can be useful for practitioners (influencers and marketers) in three possible ways. First, brands that sponsor influencers can benefit from a clear understanding of how mentions across different types of brands affect outcomes. In order to do this, we focus on one of the significant relationships identified from the Text model. Using a Ridge Regression model where each brand has a unique coefficient, we run the regressions in Step 1 and 2 again and display the results of the coefficients in Figure 14. As can be seen from the figure, the x-axis captures the attention weight directed to brand mentions and the y-axis reflects the sentiment. The brands driving the main effect are in the bottom right quadrant (positive attention weight and negative sentiment). Based on the brands in the quadrant, it appears that this relationship mainly exists for consumer electronics and video-gaming categories (about 70% of the brands in that quadrant). Thus, brands in these categories may find it useful to suggest influencers to drop brand mentions in the first 30 seconds and then test this more rigorously. They may be better off focusing on brand mentions in other parts of the video.

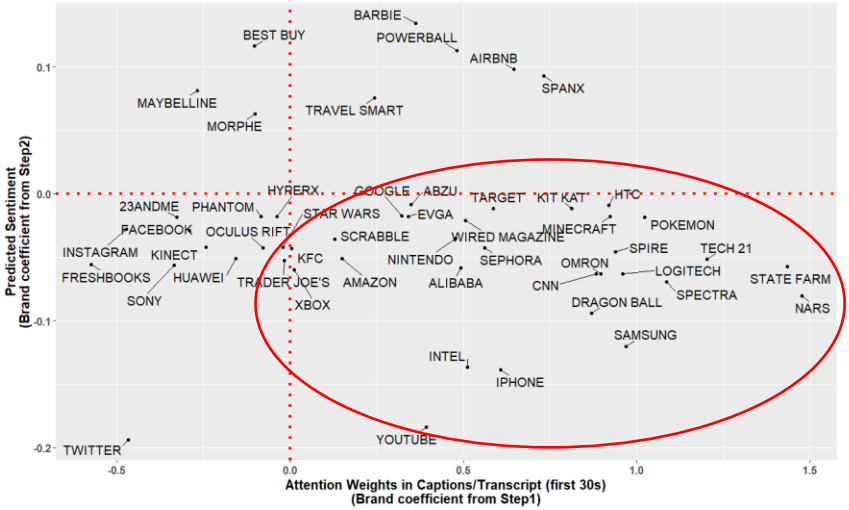

Figure 14: Brand Heterogeneity (brand mention in captions/transcript in first 30s vs sentiment)

Another possibility in terms of evaluating the effectiveness of influencer videos is to focus on the interaction between video elements (text, audio and images). However, this is non-trivial as the potential number of combinations is very large. One possible way to reduce this number and focus on relevant interactions is to draw on the findings from the literature, especially that on conventional advertising. For example, this literature has found that use of voice-over, animal images and music can reduce irritation towards the ad (Pelsmacker & Van den Bergh, 1999). Hence, we study whether interactions such as brand mention with background music, brand mention with size of person image, and brand mention with size of animal images (within first 30 seconds of video) are significantly associated with sentiment towards the video. We run corresponding regressions as done in Section 5.2 and find null effects for each of these interactions.

Second, besides evaluating specific elements or specific interactions between elements (as above), marketers may also be interested in evaluating influencer videos in a holistic manner. We develop a scoring mechanism to help them determine the impact and effectiveness of influencer videos. A video can be scored out of 100% on each of its unstructured elements when predicting any of the five outcomes. We do this with the help of the equation of the combined Ridge Regression model detailed in Section 4.4 which is reproduced below:

$$Y_{it} = g\left(X_{it}, \hat{Y}_{it_{Title}}, \hat{Y}_{it_{Description}}, \hat{Y}_{it_{Caption/Transcript}}, \hat{Y}_{it_{Audio}}, \hat{Y}_{it_{Thumbnail}}, \hat{Y}_{it_{Video\ Frames}}\right) + \epsilon_{it}$$

We begin by creating a linear Partial Dependence Plot (PDP) (J. Friedman, 2001) between $\hat{Y}_{it<unstructured\ element>}$ and Y_{it} in the training sample. We note the minimum and maximum values of $\hat{Y}_{it<unstructured\ element>}$ while predicting each of the five outcomes. We then note the values of $\hat{Y}_{it<unstructured\ element>}$ for a random video in the holdout sample. We scale it using min-max scaling to get a score out of 100% for each unstructured element while predicting each outcome. Finally, we get an overall score by weighing the score for each unstructured element with its relative importance score (based on Table 5).

As an illustration, let us take the point of view of a brand that is evaluating a particular influencer as a potential partner. As a sample, they pick this video (https://www.youtube.com/watch?v=3-oWqeA_hc4) and score it as above. From Table 11, we can see that the video's weakest performance area based on its overall score is on the engagement and popularity outcomes, with the weakest elements being captions and video frames respectively. The brands can use these scores to (a) suggest areas of improvement to the influencers, (b) compare this video to other videos from the same influencer and (c) compare this video to videos from other influencers. Similarly, the influencer can use these scores to

progressively refine their videos for the relevant outcome. Note that these summary scores are based on correlations between video elements and outcomes, so their value lies in providing directions along which improvements are most likely. In contrast, without these scores, the number of directions on which influencers and brands can work is very large.

	Views	Sentiment	Engagement	Popularity	Likeability
Title Score	83.45%	100.00%	36.60%	57.98%	75.63%
Description Score	74.76%	100.00%	34.56%	59.97%	61.16%
Captions/Transcript Score	84.17%	100.00%	6.79%	45.45%	90.60%
Audio Score	88.34%	100.00%	28.43%	47.76%	58.74%
Thumbnail Score	25.62%	100.00%	43.50%	40.33%	50.00%
Video Frame Score	90.97%	0.00%	65.17%	29.75%	43.30%
Overall Score	**77.58%**	**90.12%**	**33.24%**	**54.37%**	**69.74%**
Observed value for this YouTube video	368,796	POSITIVE	5 comments / 10K views	179 likes / 10K views	118 (likes+1)/ (dislikes+1)
Median value in dataset across 1620 videos	140,000	NA	19 comments / 10K views	220 likes / 10K views	54 (likes+1)/ (dislikes+1)

Table 11: Score for a video outside the training sample

Third, a bigger question for marketers is to understand the overall importance of branded content in these influencer videos. One way to quantify this is to look at the important elements that go into outcome predictions (as in Table 9) and decompose the variance explained by brand related content e.g., brand mentions versus other content in captions/transcript. This is illustrated in Table 12, where we show the variance explained by the presence of brand mentions in the video description to predict views, and the presence of brand mentions in captions/transcript to predict sentiment. We use the Ridge Regression model (found as best performing in Section 5.1) to measure the ability of brand mentions to predict the outcome variable and compare its performance with the Text model (BERT) that was originally used (in Section 4.1) to measure the ability of the whole text to predict the outcome variable. We find that brand mentions in description explain 7.8% of the variation in views, whereas brand mentions in captions/transcript explain 39.7% of the variation in sentiment, thus demonstrating the relatively more important role played by brand mentions in predicting sentiment towards the video.

Model	Outcome	Covariate	Holdout $\sqrt{SSE} = \sqrt{(y-\hat{y})^2}$ or accuracy	Baseline $\sqrt{SST} = \sqrt{(y-\bar{y})^2}$ or accuracy	Improve-ment over baseline	Variance explained by branded content
Ridge Regression	Views	Two indicators for brand mention in first and second half of description (first 160c)	2.2	2.3	2.4%	7.8%
BERT	Views	Description (first 160c)	1.6	2.3	30.5%	
Ridge Regression	Sentiment	Two indicators for brand mention in first and second half of captions/transcript (first 30s)	58.0%	50.0%	16.0%	39.7%
BERT	Sentiment	Captions/Transcript (first 30s)	70.2%	50.0%	40.4%	

Table 12: Variance explained by brand mentions

7. Conclusion

This paper adds to the small body of work on an important and growing marketing mechanism, influencer marketing. The main vehicle used in influencer marketing is influencer videos, with brands sponsoring and/or inserting advertising during these videos. There is virtually no research on how the elements of these videos (across text, audio and images) are related to outcomes that both influencers and marketers care about. This paper takes the first step at documenting and interpreting these relationships. Methodologically, the paper uses novel transfer learning/deep learning approaches that avoid making a tradeoff between interpretability and predictive ability. After carrying out predictions using unstructured data, interpretation is carried out ex-post by quantifying the attention paid on word-pieces in text, moments in audio and items in images while forming an association with an outcome. This information is used to find significant positive (significant) relationships between video elements and attention (gradients), followed by the determination of significant relationships between video elements and the predicted outcome of interest. An added benefit of this approach is that it allows filtering out relationships that are affected by confounding factors unassociated with an increase in attention. This significantly reduces the effort required for further causal work.

The proposed approach not only allows quantifying the relative importance of data modalities (text, audio, and images), but also allows visualization of salient regions across these modalities. This allows to provide a holistic perspective about the role of each component in predicting outcomes of interest to both influencers and brand partners. In terms of practical applications, key findings such as a brand mention (especially from consumer electronics and video game categories) in the first 30 seconds results in a *significant increase* in attention to the brand but a *significant decrease* in sentiment towards

the video, can help influencers refine their videos. Brands can also use these findings to evaluate the attractiveness of a given influencer's video by either focusing on specific elements and their interactions or analyzing them in a holistic manner for their marketing campaigns. A broader view suggests that our approach can be adapted to the analysis of (non-traditional) videos in multiple domains e.g., education and politics.

Given that the paper represents early work on this topic, it suffers from some limitations. First, as we have no access to sales data from influencer campaigns, we use proxy metrics that, while relevant to marketers, may not be perfectly correlated to business metrics. Interestingly, however, brands find it very difficult to assess the ROI of influencer marketing campaigns, suggesting that measurement of sales data is non-trivial (Bailis, 2020; Kramer, 2018). Second, as our sample includes only influencers who use brand endorsements, we cannot offer any insights about the quality of videos from those influencers who never receive such endorsements. Third, the uncovered relationships between advertising content and outcomes, while based on an increase in attention to advertising content, do not guarantee causality, and need to be validated e.g., via field experiments. Fourth, while YouTube is one of the most important influencer marketing platforms, there may be systematic differences in how influencer videos work on other channels such as Instagram or TikTok. Finally, given that different devices (e.g., mobile, desktop and tablet) can be used to access content from the same platform, identified relationships could vary by device, but our findings only capture the average effect. We hope that future work can address these limitations.

References

Aaker, D. A., & Stayman, D. M. (1990). Measuring audience perceptions of commercials and relating them to ad impact. *Journal of Advertising Research*.

Alammar, J. (2018, June 27). The Illustrated Transformer. Retrieved from http://jalammar.github.io/illustrated-transformer/

Atlas ML. (2020). Image Classification on ImageNet. Retrieved from https://paperswithcode.com/sota/image-classification-on-imagenet

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bailis, R. (2020). The State of Influencer Marketing: 10 Influencer Marketing Statistics to Inform Where You Invest. Retrieved from https://www.bigcommerce.com/blog/influencer-marketing-statistics/#what-is-influencer-marketing.

Brooks, A. (2020, January 24). As influencers increasingly create video content, what does this mean for brands? Retrieved from https://marketingtechnews.net/news/2020/jan/24/influencers-increasingly-create-video-content-what-does-mean-brands/

Burnap, A., Hauser, J. R., & Timoshenko, A. (2019). Design and Evaluation of Product Aesthetics: A Human-Machine Hybrid Approach. *Available at SSRN 3421771*.

Chakraborty, I., Kim, M., & Sudhir, K. (2019). Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Attribute Self-Selection. *Cowles Foundation Discussion Paper No. 2176, Available at SSRN: https://ssrn.com/abstract=3395012 or http://dx.doi.org/10.2139/ssrn.3395012*.

Contestabile, G. (2018). Influencer Marketing in 2018: Becoming an Efficient Marketplace. Retrieved from https://www.adweek.com/digital/giordano-contestabile-activate-by-bloglovin-guest-post-influencer-marketing-in-2018/

Cournoyer, B. (2014, March 19). YouTube SEO Best Practices: Titles and Descriptions. Retrieved from https://www.brainshark.com/ideas-blog/2014/March/youtube-seo-best-practices-titles-descriptions

Covington, P., Adams, J., & Sargin, E. (2016). *Deep neural networks for youtube recommendations.* Paper presented at the Proceedings of the 10th ACM conference on recommender systems.

Creusy, K. (2016, July 12). What is influencer marketing? Retrieved from https://www.upfluence.com/influencer-marketing/what-is-influencer-marketing

Dawley, S. (2017, May 30). Do Vanity Metrics Matter on Social Media? Yes (And No). Retrieved from https://blog.hootsuite.com/vanity-metrics/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dew, R., Ansari, A., & Toubia, O. (2019). Letting Logos Speak: Leveraging Multiview Representation Learning for Data-Driven Logo Design. *Available at SSRN 3406857*.

Dixon, C., & Baig, H. (2019, March 7). What is Youtube comment system sorting / ranking algorithm? . Retrieved from https://stackoverflow.com/questions/27781751/what-is-youtube-comment-system-sorting-ranking-algorithm

Dzyabura, D., El Kihal, S., & Ibragimov, M. (2018). Leveraging the power of images in predicting product return rates. *Available at SSRN: https://ssrn.com/abstract=3209307 or http://dx.doi.org/10.2139/ssrn.3209307*.

Dzyabura, D., & Yoganarasimhan, H. (2018). Machine learning and marketing. In *Handbook of Marketing Analytics*: Edward Elgar Publishing.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, W. (2017). Shorter-Duration TV Commercials On The Rise. Retrieved from https://www.mediapost.com/publications/article/308248/shorter-duration-tv-commercials-on-the-rise.html

FTC. (2020). Disclosures 101 for Social Media Influencers. Retrieved from https://www.ftc.gov/system/files/documents/plain-language/1001a-influencer-guide-508_1.pdf
Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., . . . Ritter, M. (2017). *Audio set: An ontology and human-labeled dataset for audio events.* Paper presented at the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM.
Google. (2020a). Add tags to videos. Retrieved from https://support.google.com/youtube/answer/146402?hl=en
Google. (2020b). How video views are counted. Retrieved from https://support.google.com/youtube/answer/2991785?hl=en
Google. (2020c). Manage ad breaks in long videos. Retrieved from https://support.google.com/youtube/answer/6175006?hl=en
Google. (2020d). Paid product placements and endorsements. Retrieved from https://support.google.com/youtube/answer/154235?hl=en
Google. (2020e). Use hashtags for video search. Retrieved from https://support.google.com/youtube/answer/6390658?hl=en
Hartmann, J., Heitmann, M., Schamp, C., & Netzer, O. (2020). The Power of Brand Selfies in Consumer-Generated Brand Imagery. *Columbia Business School Research Paper Forthcoming, Available at SSRN: https://ssrn.com/abstract=3354415 or http://dx.doi.org/10.2139/ssrn.3354415*.
Hopf, M. (2020). NLP With Google Cloud Natural Language API. Retrieved from https://www.toptal.com/machine-learning/google-nlp-tutorial
Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., . . . Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
Hu, J., Shen, L., & Sun, G. (2018). *Squeeze-and-excitation networks.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
Hughes, C., Swaminathan, V., & Brooks, G. (2019). Driving Brand Engagement Through Online Social Influencers: An Empirical Investigation of Sponsored Blogging Campaigns. *Journal of Marketing*.
Influencer Marketing Hub. (2018, October 24). 4 Factors That Affect Your YouTube Earnings Potential. Retrieved from https://influencermarketinghub.com/4-factors-affect-youtube-earnings-potential/
Influencer Marketing Hub and CreatorIQ. (2020). *The State of Influencer Marketing 2020 : Benchmark Report* Retrieved from https://influencermarketinghub.com/influencer-marketing-benchmark-report-2020/
Ismail, K. (2018, December 10). Social Media Influencers: Mega, Macro, Micro or Nano. *cmswire.com.* Retrieved from https://www.cmswire.com/digital-marketing/social-media-influencers-mega-macro-micro-or-nano/
Klear (Producer). (2019, April 8, 2020). Inlfuencer Marketing Rate Card. Retrieved from https://klear.com/KlearRateCard.pdf
Kramer, S. (2018, September 4). The Impact of Influencer Marketing on Consumers. Retrieved from https://www.themarketingscope.com/influencer-marketing-on-consumers/
Lambert, B. (2018). YouTube Ads Benchmarks for CPC, CPM, and CTR in Q4 2018. Retrieved from https://blog.adstage.io/q4-2018-youtube-benchmarks
Lanz, A., Goldenberg, J., Shapira, D., & Stahl, F. (2019). Climb or Jump: Status-Based Seeding in User-Generated Content Networks. *Journal of Marketing Research, 56*(3), 361-378.
Lee, D., Manzoor, E., & Cheng, Z. (2018). Focused Concept Miner (FCM): An Interpretable Deep Learning for Text Exploration. *Available at SSRN 3304756*.
Li, X., Shi, M., & Wang, X. S. (2019). Video mining: Measuring visual information using automatic methods. *International Journal of Research in Marketing, 36*(2), 216-231.

Liu, L., Dzyabura, D., & Mizik, N. (2018). *Visual listening in: Extracting brand image portrayed on social media.* Paper presented at the Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.

Liu, X., Lee, D., & Srinivasan, K. (2019). Large scale cross category analysis of consumer review content on sales conversion leveraging deep learning. *NET Institute Working Paper No. 16-09, Available at SSRN: https://ssrn.com/abstract=2848528 or http://dx.doi.org/10.2139/ssrn.2848528*.

Lu, S., Xiao, L., & Ding, M. (2016). A video-based automated recommender (VAR) system for garments. *Marketing Science, 35*(3), 484-510.

Maheshwari, S. (2018, November 11). Are You Ready for the Nanoinfluencers? Retrieved from https://www.nytimes.com/2018/11/11/business/media/nanoinfluencers-instagram-influencers.html

McGranaghan, M., Liaukonyte, J., & Wilbur, K. C. (2019). Watching people watch TV. *Working Paper*.

Mediakix. (2020). YouTube Sponsored Videos: Advertising & Influencer Marketing Guide. Retrieved from https://mediakix.com/influencer-marketing-resources/youtube-sponsored-videos/

Mitchell, A. A. (1986). The effect of verbal and visual components of advertisements on brand attitudes and attitude toward the advertisement. *Journal of consumer research, 13*(1), 12-24.

O'Connor, C. (2017a, April 10). Earning Power: Here's How Much Top Influencers Can Make On Instagram And YouTube. Retrieved from https://www.forbes.com/sites/clareoconnor/2017/04/10/earning-power-heres-how-much-top-influencers-can-make-on-instagram-and-youtube/#a1d07bd24db4

O'Connor, C. (2017b, September 26). Forbes Top Influencers: Meet The 30 Social Media Stars Of Fashion, Parenting And Pets (Yes, Pets). Retrieved from https://www.forbes.com/sites/clareoconnor/2017/09/26/forbes-top-influencers-fashion-pets-parenting/#1a67cea27683

Olney, T. J., Holbrook, M. B., & Batra, R. (1991). Consumer responses to advertising: The effects of ad content, emotions, and attitude toward the ad on viewing time. *Journal of consumer research, 17*(4), 440-453.

Oxford Reference. (2020). influencer. Retrieved from https://www.oxfordreference.com/view/10.1093/acref/9780191803093.001.0001/acref-9780191803093-e-630

Parsons, J. (2017, August 24). How Long Until Watching a YouTube Video Counts as a View? Retrieved from https://growtraffic.com/blog/2017/08/youtube-video-counts-view

Pelsmacker, P. D., & Van den Bergh, J. (1999). Advertising content and irritation: a study of 226 TV commercials. *Journal of international consumer marketing, 10*(4), 5-27.

Pilakal, M., & Ellis, D. (2020). YAMNet. Retrieved from https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

Rosenberg, E. (2018, October 7). How Youtube Ad Revenue Works. Retrieved from https://www.investopedia.com/articles/personal-finance/032615/how-youtube-ad-revenue-works.asp

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision, 115*(3), 211-252.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). *Mobilenetv2: Inverted residuals and linear bottlenecks.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Tabor, E. (2020, April 8). Credibility And Trust Are Key To Authentic Influencer Marketing. Retrieved from https://www.forbes.com/sites/forbesagencycouncil/2020/04/08/credibility-and-trust-are-key-to-authentic-influencer-marketing/

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.
Teixeira, T., Picard, R., & El Kaliouby, R. (2014). Why, when, and how much to entertain consumers in advertisements? A web-based facial tracking field study. *Marketing Science, 33*(6), 809-827.
Teixeira, T., Wedel, M., & Pieters, R. (2010). Moment-to-moment optimal branding in TV commercials: Preventing avoidance by pulsing. *Marketing Science, 29*(5), 783-804.
Teixeira, T., Wedel, M., & Pieters, R. (2012). Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research, 49*(2), 144-159.
Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science, 38*(1), 1-20.
Vashishth, S., Upadhyay, S., Tomar, G. S., & Faruqui, M. (2019). Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). *Attention is all you need.* Paper presented at the Advances in Neural Information Processing Systems.
Widex. (2016, August 9). The human hearing range - what can you hear? Retrieved from https://www.widex.com/en-us/blog/human-hearing-range-what-can-you-hear
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., . . . Macherey, K. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
Xie, Q., Hovy, E., Luong, M.-T., & Le, Q. V. (2019). Self-training with Noisy Student improves ImageNet classification. *arXiv preprint arXiv:1911.04252*.
YouTube. (2020). What is fair use? Retrieved from https://www.youtube.com/intl/en-GB/yt/about/copyright/fair-use/#yt-copyright-four-factors
Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). *Beyond short snippets: Deep networks for video classification.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2017). How much is an image worth? Airbnb property demand estimation leveraging large scale image analytics. *Airbnb Property Demand Estimation Leveraging Large Scale Image Analytics (May 25, 2017)*.
Zhao, K., Hu, Y., Hong, Y., & Westland, J. C. (2019). Understanding Factors that Influence User Popularity in Live Streaming Platforms. *Available at SSRN 3388949*.
Zimmerman, E. (2016, February 24). Getting YouTube Stars to Sell Your Product. Retrieved from https://www.nytimes.com/2016/02/25/business/smallbusiness/getting-youtube-stars-to-sell-your-product.html

Appendix A – BERT Encoders (in Text Model)

BERT Encoders comprise a set of 12 sequentially arranged identical encoders, and we illustrate the architecture of one encoder in Figure A1.[41] We explain an example with a sentence that has only two tokens, and this can be extended to any example that has a maximum of 512 tokens, which is the maximum limit of the pre-trained BERT model. The combined vector of the initial token embedding (x_1, x_2) and positional encoding (t_1, t_2) results in the vectors (x_1', x_2') that are passed through self-attention heads which incorporate information of other relevant tokens into the focal tokens. The architecture of the self-attention head is explained further ahead. The outputs of the self-attention head (z_1, z_2) are then added with the original input (x_1', x_2') using a residual connection (shown with a curved arrow) and normalized (using mean and variance). The outputs (z'_1, z'_2) are passed through identical feed forward networks that have a GELU (Gaussian Error Linear Unit) activation function, i.e. $gelu(x) = 0.5x\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right)$. The gelu activation combines the advantages of the ReLU (Rectified Linear Unit) non-linearity (i.e., $relu(x) = max(0, x)$) with dropout regularization. The outputs of the feed forward network are added with the inputs (z'_1, z'_2) using a residual connection and normalized again before being fed to the next encoder in sequence. In addition, each sub-layer is first followed by a dropout probability of 0.1 before being added and normalized.

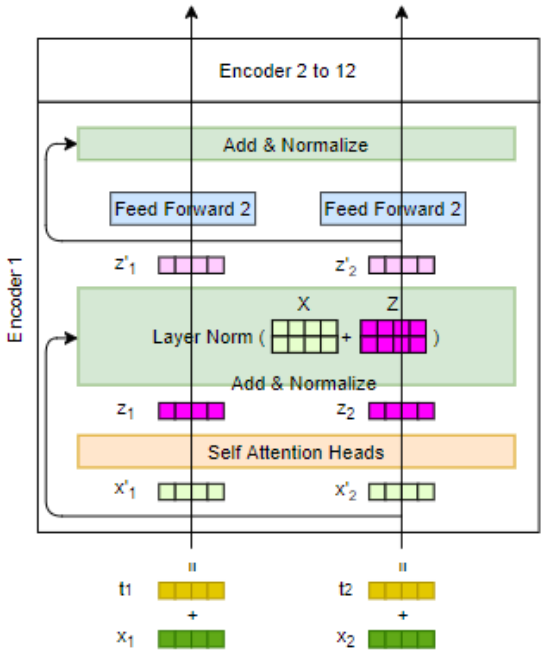

Figure A1: Encoders

[41] Our figures are inspired from the work of Jay Alammar.(see Alammar (2018) for more details)

Next, we explain the self-attention heads using the framework shown in Figure A2. There are 12 self-attention heads that capture the contextual information of each token in relation to all other tokens used in the text. In other words, this allows the model to identify and weigh all other tokens in the text that are important when learning the vector representation of the focal token. We use this to visualize the strength of association between the tokens in the text and the outcome of interest in Section 5.2.

The inputs (x_1', x_2') are concatenated and multiplied with three weight matrices, $W^q, W^k$and W^v (that are fine-tuned during model training) to get three vectors – Q (Query), K(Key) and V(Value). These three vectors are combined using an attention function (A):

$$A(Q,K,V) = z_0'' = softmax\left(\frac{Q.K^T}{\sqrt{d_k}}\right).V$$

where, d_k, the dimension of the Key vector, is chosen to be 64 and is equal to the dimensions of the other two vectors d_q and d_v; and $softmax(x) = \frac{e^{x_i}}{\sum_{i=1}^{m} e^{x_i}}$. The division by $\sqrt{d_k}$ is performed to ensure stable gradients. The computation of z_0'' is for one attention head, and this is carried out in parallel for 11 additional attention heads to give us 12 vectors, $z_0'' \ldots z_{12}''$, which are concatenated to produce z''. This is multiplied with a weight vector W^O (which is fine-tuned during model training) to produce output (z_1, z_2). The use of 11 additional attention heads allows the model to capture more complex contextual information.

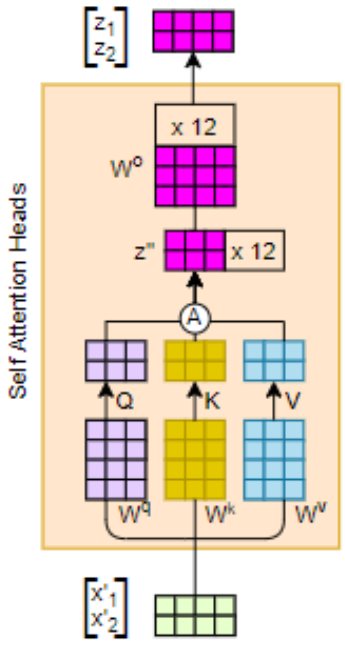

Figure A2: Self-Attention Heads

Appendix B – MobileNet V1 followed by Bi-LSTM with Attention (in Audio Model)

The MobileNet v1 architecture is illustrated in detail in Table B1. Each row describes Stage i with input dimension $[\widehat{H}_i, \widehat{W}_i]$ (resolution), output channels $\hat{C}_i$ (width) and $\hat{L}_i$ layers (depth).

Stage i	**Operator** $\hat{F}_i$	**Resolution** ($\hat{H}_i$ x $\widehat{W}_i$) (Height x Width)	**Width** $\hat{C}_i$ (Channels)	**Depth** $\hat{L}_i$ (Layers)	**Pre-trained Weights**
1	Conv, k3x3, s2	96 x 64	32	1	Yes
2	MConv, k3x3, s1	48 x 32	64	1	
3	MConv, k3x3, s2	48 x 32	64	1	
4	MConv, k3x3, s1	24 x 16	128	1	
5	MConv, k3x3, s2	24 x 16	128	1	
6	MConv, k3x3, s1	12 x 8	256	1	
7	MConv, k3x3, s2	12 x 8	256	1	
8	MConv, k3x3, s1	6 x 4	512	5	
9	MConv, k3x3, s2	6 x 4	512	1	
10	MConv, k3x3, s2	3 x 2	1024	1	
11	Global Average Pooling	3 x 2	1024	1	
12	Dense	1 x 1	521	1	

Table B1: MobileNet-v1 architecture

Stage 1 has a regular convolution operation, whereas Stage 2 to 10 have the Mobile Convolution which is the main building block of the architecture. It is represented as "MConv, k x k, s" where k x k = 3 x 3 is the size of the kernel and $s = \{1,2\}$ is the stride. MConv divides the regular convolution operation into two steps – depth wise separable convolutions and point wise convolution, thus increasing the speed of computation (see Howard et al. (2017) for details). Stage 11 has a Global Average Pooling Layer that averages the inputs along its height and width and passes its output to Stage 10 which is a Dense output layer with 521 logistic functions that gives the per class probability score corresponding to the 960 ms input segment. We use a hop size of 490 ms so that we get an even number of 60 time step predictions corresponding to the 30 seconds of input. The resulting output vector has a dimension of 521x60 (audio classes x time steps) for each 30 second clip.

The output from MobileNet v1 is passed as input to the Bi-LSTM with attention mechanism, shown in Figure B1.We use two layers of LSTM cells – the first layer is a 32 unit Bidirectional LSTM layer and the second layer is a 64 unit (unidirectional) LSTM layer. They are separated by an attention mechanism as shown in Figure B1. Each audio segment x_m <521,1>, where m is the total number of moments (time steps), is passed as input to each cell of the Bidirectional LSTM layer. This layer is made bidirectional to allow it to capture the interdependence between sequential audio segments from both

directions. The sequential nature of LSTM cells in a layer allow the model to capture dependencies between audio segments that are separated from each other (see the LSTM paper by Gers et al. (1999) for more details). We adopt the attention mechanism used for neural machine translation by Bahdanau et al. (2014) to help the Bi-LSTM model focus on more important parts of the input. The mechanism weighs the output activations ($a^{<t>} = [\vec{a}^{<t>}, \overleftarrow{a}^{<t>}], t = 1 \text{ to } m$) from each cell of the pre-attention Bi-LSTM layer before passing the contextual output, $c^{<t>}$, to the post-attention LSTM layer above it. In addition, each cell of the attention mechanism takes as input the output activation $s(t-1)$ from each preceding cell of the post-attention LSTM layer which allows it to factor in the cumulative information learnt by the model till that time step (see Bahdanau et al. (2014) for more details on the attention mechanism). The output of the last cell in the post-attention LSTM layer is passed to an output layer which has a linear activation function for the four continuous outcomes and a sigmoid activation function for sentiment. The context vector $c^{<m>}$ from the last cell of the attention mechanism allows visualization of the relative weights placed by the model along the time dimension of the input in order to form an association with the outcome of interest. Audio moments that have higher weight are more important while forming an association between the audio clip and the outcome of interest.

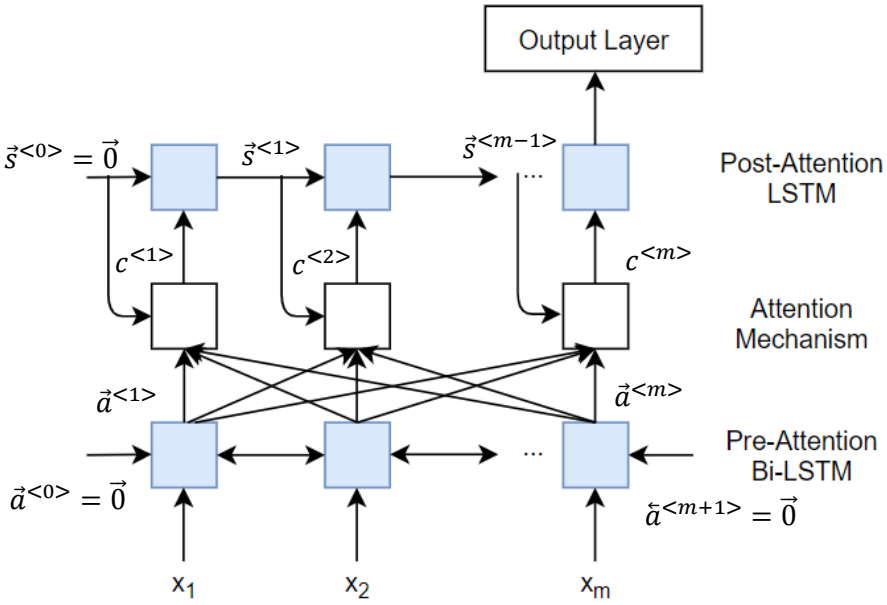

Figure B1: Bi-LSTM with Attention

Appendix C – EfficientNet-B7 Architecture and Combination Architectures (in Image Model)

The architecture of EfficientNet-B7 customized to our input dimension of 270x480x3 (where 3 corresponds to the pixel intensities for Red, Green and Blue channels) is shown in Table C1. Each row describes Stage i with input dimension $[\widehat{H}_i, \widehat{W}_i]$ (resolution), output channels $\hat{C}_i$ (width) and $\hat{L}_i$ layers (depth). B7 is the model with the highest uniform increase in resolution, width and depth of the model as compared to a baseline model B0 used by Tan and Le (2019). Scaling uniformly across the three dimensions (i.e. compound scaling) allows the model to better capture salient regions in images (see Tan and Le (2019) for details).

Stage i	**Operator** $\hat{F}_i$	**Resolution** ($\widehat{H}_i$ x $\widehat{W}_i$) (Height x Width)	**Width** $\hat{C}_i$ (Channels)	**Depth** $\hat{L}_i$ (Layers)	**Pre-trained Weights**
1	Conv k3x3, s2	270 x 480	64	1	Yes
2	MIBConv, e1, k3x3, s1	135 x 240	32	4	
3	MIBConv, e6, k5x5, s2	135 x 240	48	7	
4	MIBConv, e6, k5x5, s2	68 x 120	80	7	
5	MIBConv, e6, k3x3, s2	34 x 60	160	10	
6	MIBConv, e6, k5x5, s1	17 x 30	224	10	
7	MIBConv, e6, k5x5, s2	17 x 30	384	13	
8	MIBConv, e6, k3x3, s1	9 x 15	640	4	
9	Global Average Pooling	9 x 15	2560	1	No
10	Dense	1 x 1	1	1	

Table C1: EfficientNet-B7 architecture

Stage 1 has the regular convolution operation, whereas Stages 2 to 8 comprise the main building block of the architecture which is the Mobile Inverted Bottleneck Convolution, "MIBConv, e, k x k, s" where $e = \{1,6\}$ is the expansion factor, k x $k = \{3x3, 5x5\}$ is the size of the kernel and $s = \{1,2\}$ is the stride. The strength of MIBConv lies in its ability to identify important features that are encoded in lower dimensional subspaces of images (see Sandler et al. (2018) for details). Furthermore, each MIBConv block is followed by a squeeze-and-excitation network that provides a weighted average to each channel output instead of a simple average, thus improving model performance (see Hu et al. (2018) for details).

To analyze thumbnails, we use the pre-trained weights from Stage 1 to 8, and tune the weights of Stage 9 and 10. Stage 9 has a Global Average Pooling Layer that averages the inputs along its height and width and passes its output to Stage 10 which is a Dense output layer. The output layer has a linear activation function for the four continuous outcomes and a softmax activation function for sentiment.

To analyze video frames, we compare the performance of four 'combination architectures' shown in Figure C. Figure C1 shows the Bi-LSTM approach that captures sequential information from different video frames. Each EfficientNet-B7 model takes a different video frame as input and provides the output from Stage 8 to the Global Average Pooling (GAP) Layer. This is followed by Dense Middle Layers (that use ReLU activation for continuous outcomes and sigmoid activation for the binary outcome), which is followed by a single Bi-LSTM layer with 256 memory cells, and finally a Dense output layer (that uses linear activation for continuous outcomes and softmax activation for the binary outcome). Figures C2, C3 and C4 show three approaches that preserve the spatial information across different video frames. The Max-GAP approach finds the maximum value across the [9x15x2560] Stage 8 output from each EfficientNet-B7 model, which is followed by a GAP layer that reduces the dimensions to [1x1x2560]. The GAP-Max approach first finds the global average across each Stage 8 output, which is then followed by the Max operation, while the C-GAP approach concatenates the GAP outputs. Across all four approaches, we use the pre-trained weights from Stage 1 to 8 for each EfficientNet-B7 model and tune the weights of the final layers.

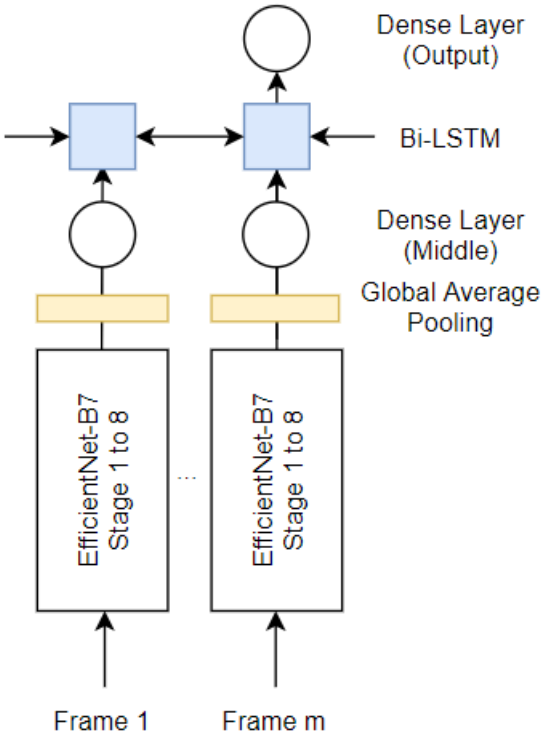

Figure C1: Bi-LSTM

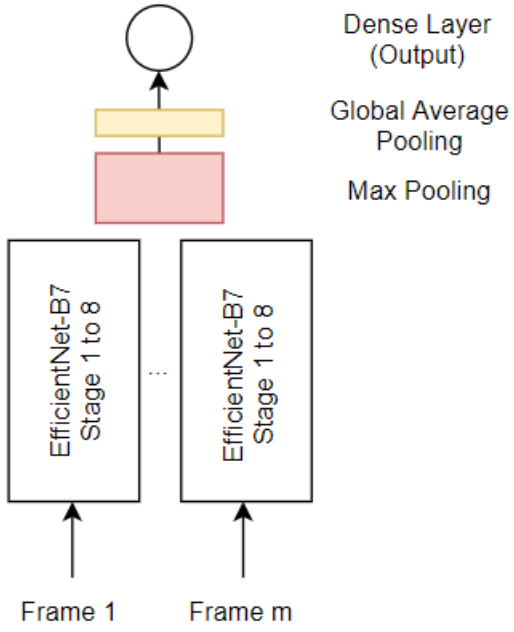

Figure C2: Max-GAP

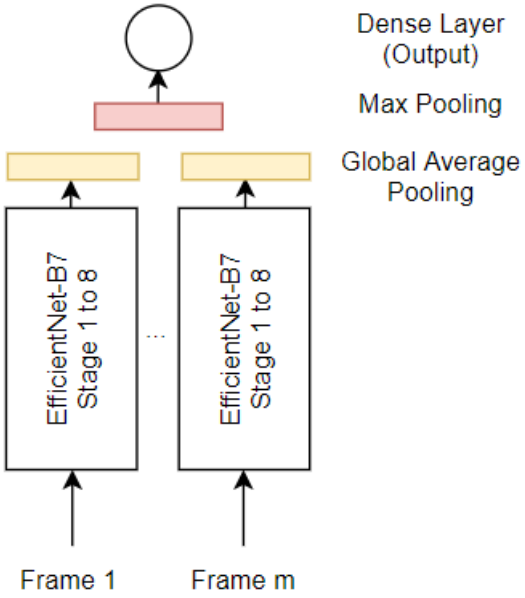

Figure C3: GAP-Max

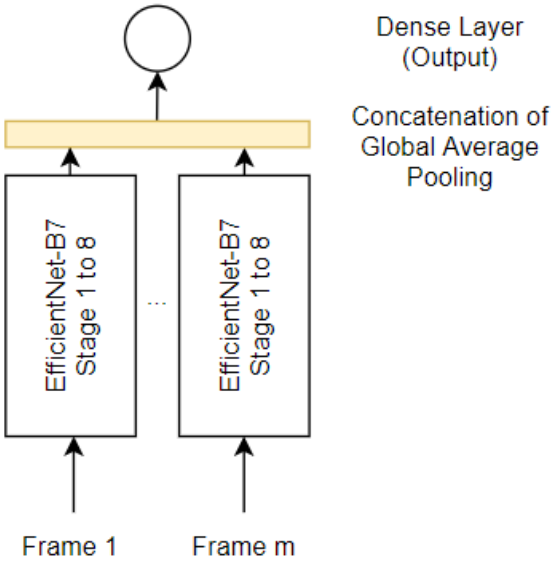

Figure C4: C-GAP

Appendix D – Comparison of Model Performance with Benchmarks

We first compare the predictive performance from the BERT Text Model with four benchmarks in Table D1. The benchmark models include an LSTM (with a 300 dimensional Glove word vector embedding), CNN model (X. Liu et al., 2019), CNN-LSTM (Chakraborty et al., 2019) and CNN-Bi-LSTM.

Outcome	Covariate	LSTM	CNN	CNN-LSTM	CNN-Bi-LSTM	BERT
Views	Title	2.11	1.73	1.75	1.68	1.66
	Description (first 160c)	2.27	1.69	1.68	1.67	1.57
	Captions/transcript (first 30s)	2.28	1.82	1.80	1.76	1.75
Sentiment	Title	0.67	0.70	0.70	0.70	0.72
	Description (first 160c)	0.50	0.69	0.69	0.69	0.69
	Captions/transcript (first 30s)	0.67	0.70	0.70	0.70	0.70

Table D1: Comparison of Text Model performance in holdout sample
(RMSE for Views; Accuracy for Sentiment)

As can be seen in Table D1, BERT has the best performance for both views (lowest RMSE) and sentiment (highest accuracy), with a maximum performance improvement of 6% for 'views-description' as compared to CNN Bi-LSTM.

We then compare the model performance of the Audio model as per the benchmarks discussed in Section 4.2 and present the results in Table D2. We find that the addition of MobileNet v1 (Model 2) helps improve accuracy when predicting sentiment, but there is no performance improvement when predicting views. Addition of the attention mechanism (Model 3) results in an improvement in both RMSE (by 10%) when predicting views and accuracy (by 1.5%) when predicting sentiment, thus demonstrating the benefit of capturing relative attention weights in the model.

Outcome	Model 1: Mel Spectrogram + Bi-LSTM	Model 2: Mel Spectrogram + MobileNet v1 + Bi-LSTM	Model 3: Mel Spectrogram + MobileNet v1 + Bi-LSTM + Attention
Views	2.19	2.19	1.97
Sentiment	0.59	0.64	0.65

Table D2: Comparison of Audio Model performance in holdout sample
(RMSE for Views; Accuracy for Sentiment)

Next, we compare the performance of the (pre-trained) EfficientNet-B7 with a 4-layer CNN model using thumbnails in Table D3. We see a substantial improvement in both RMSE (by 41%) and accuracy (by 26%) when using EfficientNet-B7, thus demonstrating the benefits of both deeper architecture and transfer learning with image data.[42]

Outcome	CNN	EfficientNet-B7
Views	5.20	3.09
Sentiment	0.54	0.68

Table D3: Comparison of Thumbnail Model performance in holdout sample (RMSE for Views; Accuracy for Sentiment)

Next, we compare the performance of the four Video Frame Models in Table D4 using two frames in each video clip – 0 sec and 30 sec.

Outcome	Bi-LSTM	Max-GAP	GAP-Max	C-GAP
Views	2.28	3.16	2.88	5.53
Sentiment	0.66	0.66	0.66	0.66

Table D4: Comparison of Video Frame Model (0s,30s) performance in holdout sample (RMSE for Views; Accuracy for Sentiment)

We find that the Bi-directional LSTM architecture which captures the sequential information from two video frames performs better than the other three models that capture only spatial information while predicting views. However, all four models perform equally well in predicting sentiment. This demonstrates that capturing sequential information is more important for predicting views but not as important for predicting sentiment. As the Bi-LSTM model is the best overall, we use it to predict all the outcomes. Furthermore, we demonstrate sequential improvement in predictive performance when we add additional video frames to the model by reducing the time interval by half at each step, in Table D5.

Time Interval	Covariate	Views	Sentiment
30 s	Video Frames (0s, 30s)	2.28	0.66
15 s	Video Frames (0s, 15s, 30s)	2.23	0.67
7.5 s	Video Frames (0s, 7.5s, 15s, 22.5s, 30s)	2.23	0.68

Table D5: Bi-LSTM Video Frame Model (with different time intervals) (RMSE for Views; Accuracy for Sentiment)

[42] Note that we are unable to tune an entire EfficientNet-B7 (without transfer learning) to demonstrate only the incremental benefit of transfer learning because of computational constraints that can be achieved at a low cost.

We find that using an additional frame at 15 sec helps improve prediction of views and sentiment. Adding two more frames at 7.5 sec and 22.5 sec improves prediction of sentiment but does not result in improved prediction of views.[43] The use of five frames results in overall best performance for both outcomes.

Last, we show the performance of the Combined Model from Section 4.4 on the holdout sample in Table D6. We use four linear models – OLS, Ridge Regression (L2 penalization), LASSO (L1 penalization), Elastic Net (0.5L1 and 0.5L2 penalization), and three non-linear models – Deep Neural Net (with three hidden layers), Random Forests and Extreme Gradient Boosting (XGBoost). We find that Ridge Regression has the best performance on the holdout sample for all the continuous outcomes (lowest RMSE) and also the binary outcome (highest accuracy).

	Views	Sentiment	Engagement	Popularity	Likeability
OLS	1.46	0.74	0.80	0.64	0.78
Ridge Regression	1.11	0.75	0.73	0.56	0.72
LASSO	2.26	0.73	0.97	0.80	1.02
Elastic Net	2.26	0.74	0.97	0.80	1.02
Deep Neural Net	1.13	0.75	0.76	0.59	0.73
Random Forests	1.23	0.74	0.74	0.58	0.75
XGBoost	1.44	0.72	0.81	0.64	0.78

Table D6: Performance of different Combined Models on holdout sample
(RMSE for Views, Engagement, Popularity and Likeability; Accuracy for Sentiment)

[43] We are unable to test performance with smaller time intervals due to limitations on computational performance that can be achieved at a low cost.

Appendix E: Interpreting Interaction Effects in Text Model

We study interaction effects in Step 1 to answer the following question:

1) Brand Attention – Brand Proportion and Brand Position: Does attention change based on whether the brand is part of a longer text (proportion) and the brand position in text?

$$\log(AttentionWeight_{itj}) = \alpha_i + \gamma X_{it} + \beta_1(BIT_{itj}) + \beta_2(TP_{itj}) + \beta_3(LOTX_{itj}) + \beta_4(BIT_{itj} * LOTX_{it}) + \beta_5(BIT_{itj} * TP_{itj}) + \epsilon_{itj} \quad \text{E1}$$

In Step 2, we answer the following questions related to interaction:

2a) Brand Presence - Brand Proportion: Is there an interaction effect between brand presence in text and overall length of text?

$$PredictedOutcome_{it} = \alpha_i + \gamma X_{it} + \beta_1(BITX_{it}) + \beta_2(LOTX_{it}) + \beta_{3_{it}}(BITX_{it} * LOTX_{it}) + \epsilon_{it} \quad \text{E2}$$

2b) Brand Presence - Lead or End with Brand: Is brand presence in first or second half of each text associated with predicted outcome?

$$PredictedOutcome_{it} = \alpha_i + \gamma X_{it} + \beta_1(BIFTX_{it}) + \beta_2(BISTX_{it}) + \beta_3(LOTX_{it}) + \epsilon_{it} \quad \text{E3}$$

where, $BIFTX$ & $BISTX$ are Brand Indicators in First half of each Text and Second half of each Text, respectively.

The values of the coefficients of interest in each of the above equations are shown in Table E1. The values in the table reflect a percent change in the non-log-transformed outcome (e.g., views and not log(views)) when a covariate is present or increases by one unit.[44] We do not find significant evidence at the intersection of Step 1 and Step 2 to show that the effect of brand mentions can vary based on length of text or its position in the text.

[44] LOTX (Length of Text) has been mean centered to allow for interpretability of the coefficient.

		Step 1 - Eq(E1)			Step 2 - Eq(E2)		Step 2 - Eq(E3)	
Model for	Data Type	Covariate						
		BIT	BIT x LOTX	BIT x TP	BITX	BITX x LOTX	BIFTX	BISTX
Views	Title	29.92*	2.84*	-2.84W	0.34	-2.22	3.21	-19.60
	Desc	8.77	1.35*	-0.02	54.13*	1.28	29.86W	46.89*
	Tran	4.29	0.41*	-0.23	6.26	0.05	25.31	-7.32
Sentiment	Title	-18.37	-1.8	-0.5	-12.13	2.32	-44.25	104.53
	Desc	26.02	1.52*	-2.42*	-72.34W	11.40*	-18.95	-41.99
	Tran	40.08W	0.22	-0.20	-51.98	-4.49W	-50.78	-74.99
Engagement	Title	27.87*	-0.24	0.98	-4.62	-0.71	-12.09	1.50
	Desc	-9.64	1.95*	-0.19	4.03	0.15	7.00	-0.64
	Tran	461.32*	0.81*	-0.05	15.86W	-0.34W	-0.32	6.57
Popularity	Title	22.32W	0.76	-2.03	-9.33	-0.27	-11.40	-13.8
	Desc	85.75*	-0.76	-1.53*	2.36	0.85	11.40	-1.29
	Tran	107.04*	0.76*	-0.11	-2.41	0.42*	8.75	8.79
Likeability	Title	42.59*	0.53	-0.92	-17.27	2.57	-9.20	-4.40
	Desc	130.50*	-0.61	-0.7W	10.35	1.52*	15.58	2.50
	Tran	87.47*	1.19*	-0.44	0.87	0.17	9.37	5.15

Table E1: Results of the Text Regression Models with interactions
(– Significant ($p < 0.05$); W – Weakly Significant ($0.05 \leq p < 0.1$))*

Appendix F: Interpreting Results for Middle and End of Videos

We interpret the results of the deep learning models on the middle 30 sec and last 30 sec of videos as done in Section 5.2. We focus on all the outcomes except views because, as mentioned in Section 3.3, view count is determined by viewing 30 seconds of a video. This is more likely to be the first 30 seconds unless viewers immediately skip to other locations in the video for which we have no demonstrated evidence of it happening on YouTube. We find eight significant results for the Audio model, but no significant results for the Image and Text model. The results are shown in Table F1.

	Significant *increase* in attention (A) and significant *increase* in outcome (O)	Significant *increase* in attention (A) but significant *decrease* in outcome (O)
Outcome	Audio Model	Audio Model
Sentiment	more speech (without simultaneous music) in *middle* 30 sec of audio A: 2.88% O: 10.80% or more music (without simultaneous speech) in *middle* 30 sec of audio A: 3.10% O: 24.65%	
Engagement		more music (without simultaneous speech) in *last* 30 sec of audio A: 2.95% O: - 1.05%
Popularity	more speech (without simultaneous music) in *last* 30 sec of audio A: 5.60% O: 0.07%	more speech (without simultaneous music) in *middle* 30 sec of audio A: 0.98% O: - 0.21% or more music (without simultaneous speech) in *middle* 30 sec of audio A: 1.05% O: - 0.77%
Likeability	more speech (without simultaneous music) in *last* 30 sec of audio A: 3.63% O: 0.26% or more animal sounds in *last* 30 sec of audio A: 6.60% O: 1.02%	

Table F1: Results from interpreting Regression Models on middle and end of video

Overall, the results for the middle and end of audio are qualitatively similar with the results for the beginning of audio with some differences (e.g., more speech without simultaneous music in the last 30 sec of audio is associated with an increase in popularity which was not found to be true for the first 30 sec of audio).