Background Check:
Rating Zoom backgrounds with text and image classification

Wen Si, Isha Hameed, Parth Dhyani, Drishaan Jain, Naitian Zhou
University of Michigan
{wsi,ishameed,pdhyani,drishaan,naitian}@umich.edu

Abstract

The COVID-19 pandemic has forced many communications to enter the virtual setting; video calls have become the new norm for meetings, interviews and even social gatherings. This has allowed an unprecedented view into everyone's homes, which means the background of one's video becomes just as important a part of how one presents oneself as fashion choices, hair styles and language. In this project, we use ratings of video backgrounds generated by public Twitter account, @ratemyskyperoom, to determine what makes for good or bad video backgrounds.

1 Project Description

Since the beginning of the lockdown for COVID-19 in April of 2020, Twitter user @ratemyskyperoom has tweeted thousands of ratings for video conference backgrounds from various news channel interviews. These tweets contain descriptions of the backgrounds as well as commentary on the person in the frame and an attached photo of the background, which ultimately all funnel into a single rating out of 10. As of March 28th, 2021, @ratemyskyperoom is incredibly successful, with a 391,400 followers and a "Room Rater Shop" online store, where patrons can purchase anything from vases to T-shirts with the account's logo. The account has served as comedic commentary on elements of work from home backgrounds, but also can serve as insight as to what makes a good video conferencing background. As video conferencing has become the primary method of communication during the COVID-19 pandemic and will continue to be a important tool even past the pandemic, this findings from this analysis could prove a useful aid for online commmunications.

Figure 1: Example @ratemyskyperoom tweet

The primary goal of this project is to analyze what makes a background "good" and what makes a background "bad" from their tweet. These ratings can be impacted by attributes of background, such as lighting, color, objects in the background, video quality, or by the user's opinion of the person in the video conference. From there, we can build a classifier that can predict from the words in the tweet whether the background will be rated above the median rating or below the median rating. Further, we use computer vision as a supplement to textual analysis to find what makes a background "good" or "bad". The classifier would use the background image that the tweet linked in addition to the actual tweet text it-

self to classify a tweet as above the median rating or below the median rating.

We find that, though the @ratemyskyperoom account heavily favors identity-based characteristics, more spartan backgrounds tend not to perform as well compared to busier, more visually stimulating ones.

2 Related Work

2.1 Video Backgrounds in Context

In day-to-day social interactions, a necessary effort is spent on self-presentation; this is the communicative work done to project an identity and situate oneself in an interactional context (Goffman and others, 1978). This remains true in online communications, where users use a variety of mechanisms to project their digital persona: avatars, photos, videos, friends lists, biographies, etc (Bullingham and Vasconcelos, 2013). Video conferencing backgrounds are no different from other digital expressions of self. Communications may be hampered by increased self-consciousness and hamper effective communication (de Vasconcelos Filho et al., 2009; Miller et al., 2017)

2.2 Text Classification

Kouloumpis, Wilson, and Moore (2011) performed a Twitter sentiment analysis, training on a corpus of about 4 million tweets. Their study utilized the hashtags in the tweets to identify whether the tweet was positive, negative, or neutral. Each of the 1,000 most used hashtags in the set were analyzed to sort the tags that were the strongest indication of polarity. Similar to EECS 486 assignments, the study utilized unigram and bigram basesline classification after preprocessing to remove stopwords and normalize terms. The method also applied some Twitter-specific processing, including rewriting abbreviations and utilizing capitalized words and repeated punctuation as indicators of emotion. The results of the study showed that *n*-grams had the best evaluation performance, over specific lexicon and microblogging features.

Goel, Gautam, and Kumar (2016) use the Naive Bayes method to perform a Twitter sentiment analysis. Their algorithm closely aligns with our Naive Bayes approach, as their study focuses around ranking tweets as "positive" or "negative" based on the individual words. The team first preprocesses data by removing URLs, usernames, slang, and symbols, similar to our data cleaning. The study was based on a training corpus size of 16 million tweets, with more recent tweets extracted for sentiment analysis. The team also utilized the SentiWordNet library that classifies individual words with a positivity score, negativity score, and objectivity score, in tandem with the Naive Bayes method. A major improvement in the study was based on the classification of emoticons in tweets, as these were strong indicators of sentiment. Overall, the team's Naive Bayes classifier had an efficiency of 58.40%.

2.3 Image Classification

Zhang and Jatowt (2019) investigated the effectiveness of CNN (Convolutional Neural Network) models in predicting the popularity (measured in terms of number of retweets and number of likes) of tweets containing images. Specifically, their study included designing a CNN that could extract local image features, as well as allow an additional information input to be used for training alongside the images, such as the number of followers for the "tweeter" and hours elapsed from time of posting. This CNN was then trained on a collection of around 110,000 tweets, each of which have accumulated more than 100 retweets. The results of the study found that the hybrid CNN was able to make better predictions than pre-trained networks built for object detection, and were comparable with state-of-art text-based methods, specifically in predicting relative popularity.

Real estate agents catalogue large amounts of photos into groups in order to pick the most effective ones for selling the property. Bappy et al. (2017) investigated using a deep neural network to classify real estate images into room categories, like bathrooms and bedrooms, in addition to classifying different types of flooring or countertops with the goal to help real estate agents save time. Their process involved first enhancing images from real-estate websites to remove shadows through applying contrast-limited adaptive histogram equalization (CLAHE), and then classifying the images through their model, built in Tensorflow. Bappy, Barr, Srinivasan, and Roy-Chowdhury claim to be the first to have used

rating	cleaned_tweet	original_tweet	img_url
9	new view love the...	New view. Love the...	https://pbs.twmg...

Table 1: Schema of processed data table of tweets extracted from the account

computer vision on real estate images, and have published their datasets to encourage further research in the area.

3 Data Collection and Annotation

3.1 Collecting Tweets

The `twint` software library was used to scrape all tweets from the @ratemyskyperoom Twitter account. This returned a total of 19,353 tweets as of March 17, 2021. These tweets include ratings, but also product promotions, retweets and replies. Here is an example rating tweet:

> Nice inset shelf set up. Pro bracket. Icons are points hang on right. Reframe. 8/10

The tweet makes some comments about the decor, then assigns a numerical rating. To remove irrelevant tweets, we first checked that the tweet was not a reply to another user, then used a regular expression to see if a numerical rating out of 10 is present.

3.2 Automatic Labeling and Preprocessing

We also used the regular expression to automatically extract the rating from the tweet. Because of the structured format of tweets from the @ratemyskyperoom account, we were able to accomplish the rating labeling automatically. After extracting ratings, we process the tweet to remove the rating and the link to the tweet, which the Twitter API automatically appends. To further preprocess the data, we used the following procedure:

1. Remove any tweets which are replies or tweets which do not contain a rating
2. Preprocess words by lowercasing
3. Remove punctuation, hashtags and usernames

After processing tweets, there are 12,619 relevant tweets in the dataset. Table 1 shows an example from the dataset.

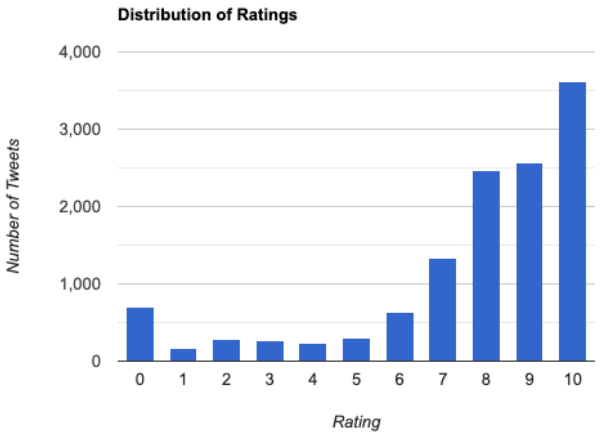

Figure 2: Ratings are skewed left, with some negatives and many zeros.

3.3 Data Exploration

Ratings were all integers out of ten. While negative ratings exist, no rooms scored higher than 10. In Figure 2, we show the distribution of ratings (out of 10). There are not many rooms with ratings between 0 and 6, so the distribution is more skewed towards the two ends of the rating scale. Most rooms receive high ratings and the probability of a low rating is much lower, but there are disproportionately more zero ratings compared to other low ratings.

In addition, after removing stopwords, we noticed a difference between the most common words in rooms rated zero and those rated as 10. The most common words for rooms rated 10 were: art, love, great, plant, and lighting, while the most common words in rooms rated 0 were: update, hostage, not, video, and still. It appears that more adversarial words were used in lower-rated rooms than higher-rated rooms.

Because tweets are limited to 250 characters, we hypothesized that the placement of certain words would be highly correlated with the placement of other words. We conducted the same frequency analysis for bigrams of the tweets, which showed that some notable common bigrams for rooms rated 10 were: (historic, skype), (psa, canada), (happy, birthday), (well, lit), and (love, art). Some notable common bigrams for rooms rated 0 were: (hostage, video), (rotten, pineapple), (pineapple, update), (dumb, dumb), (politician, update), (suntanned, politician), and (conspiracy, theories). From

these bigrams, we hypothesized that many of the ratings would be more correlated with the context of the background and the occasion, rather than actual visual elements of the background as many of the bigrams are more related to opinion or occasion, such as "happy birthday" and "dumb dumb".

Upon further analysis, the tweets appear to be heavily influenced by the user's political ideology. The phrase "rotten pineapple update" is associated with tweets related to politicians that the user criticizes, and the phrase "suntanned politican update" is associated with tweets about politicans who traveled to other countries during the pandemic. The user also appears to be Canadian or heavily influenced by Canadian politics, as many of the tweets are about Canadian politicians and Canada. Below are some examples of these tweets:

> Happy Birthday. @HillaryClinton was born this day in 1947. Always a 10/10 with us.
>
> Suntanned Politician Update. @TracyAllardUCP decided a traditional family vacation in Hawaii was more important than health orders. 0/10 #cdnpoli #ableg
>
> Rotten Pineapple Update. Sunday mornings are difficult for Dumb Dumb. 0/10 @DonaldJTrumpJr

Tweets that rated rooms lowest or highest (with a 0 or a 10) tended to have the highest user engagement, measured by likes and retweets. Ratings of 0/10 appear most popular by these metrics. The trend in engagement is portrayed in Figures 3 and 4, which plot the mean number of likes and retweets for each subset of tweets by rating.

Other statistics about our data may be found in Table 2.

4 Text Classification

4.1 Methods

We attempted two tasks: first, we performed binary classification on whether a given tweet will receive a "good" or "bad" rating, defined as scoring above ($>$) or below ($\leq$) the median rating (8/10). We trained on 80% of the dataset, holding 10% for validation

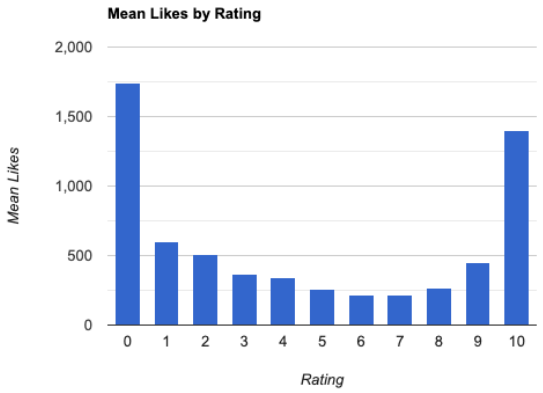

Figure 3: The highest and lowest ratings tend to get the most likes.

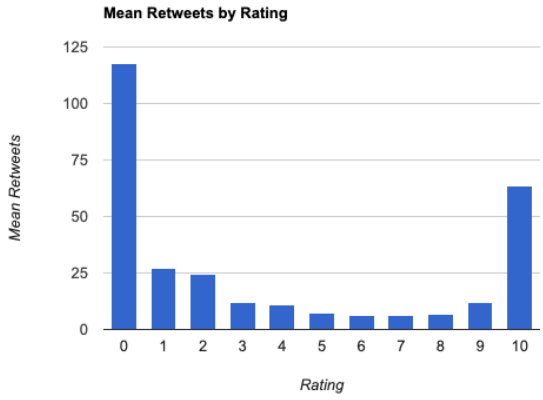

Figure 4: The highest and lowest ratings tend to get the most retweets, with the lowest rankings being most popular.

Count	12619.0
Mean	7.625
Std. Deviation	2.940
Minimum	-20.0
25%	8.0
25%	7.0
50%	8.0
75%	10.0
Maximum	100.0

Table 2: Basic statistics of the processed data

and 10% for testing. The validation set was used for hyperparameter tuning.

As features to the model, we first include the tweets in a bag-of-words representation.

Because we would like to analyze what make for good rooms, we chose simpler, more interpretable models. Specific tweet content, including words that reference light fixtures and plant decor will indicate "good" or "bad" ratings. We plan to compare the performance of a support vector machine (with linear kernel), logistic regression, naive Bayes and random forest models.

Finally, we preprocessed our data differently to see if that would have a significant impact on our models. We first removed common stopwords from the tweets and then we converted our tweets into bigrams. In both cases, we ran naive Bayes and logistic regression.

4.2 Evaluation

We evaluate all models on the 10% slice of our dataset reserved for testing. For the classification evaluation, we compare the AUROC, F1-scores, and accuracy.

5 Topic Modeling

In the nature of discerning general patterns between the tokens within a tweet and the rating assigned to a tweet, we decided to investigate the possibility of different "topics" providing some indication for the rating of a tweet. In other words, we wanted to see whether a tweet belonging to a particular topic would indicate a certain rating for the tweet. In order to accomplish this, we ran Latent Dirichlet Allocation (LDA) topic modeling on the text with both 5 topics and 10 topics. LDA topic modeling utilizes various statistical models to extract abstract topics from a given text. Once the data with each preprocessing step was fit into an LDA model for each number of topics, the resulting topic model was applied back to each tweet, outputting a vector outlining the likelihood/probability of each tweet belonging to a certain topic. This vector of topic probabilities was then appended as features for logistic regression, which was then evaluated with AUROC, F1, and accuracy.

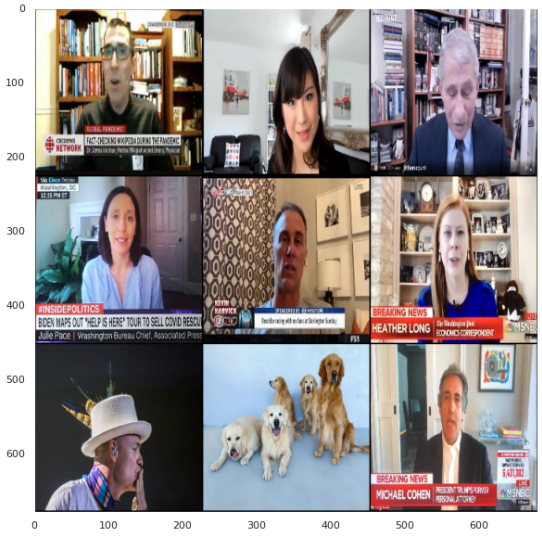

Figure 5: Sample images for image classification

6 Image Classification

Room rating tweets also usually include at least one image of the room in question. We wanted to see if there was useful information to extract from the images, so we also train a binary image classifier, using the same labels as the binary text classification task.

6.1 Data

We downloaded images corresponding to rating tweets. Of 12,619 rating tweets, 11,858 had associated images, which we downloaded. Images were further preprocessed to have dimension of 224×224 pixels. See Figure 5 for examples of images from rating tweets after preprocessing.

6.2 Training

The data was split into 80% train, 10% validation, and 10% test data. Prior to training, we also normalized images to have mean of zero and standard deviation of one, using the mean and standard deviation of the validation split.

We performed fine-tuning on a pretrained ResNet-18 model. We trained the model with a cross-entropy loss objective function for 7 epochs. However, from the training curve, we found the model yielded good results after the first epoch, and used the model after training for 4 epochs for evaluation.

7 Results

We achieve reasonable results for our classifiers (see Table 3 and Section 8 for details). We use these models to analyze which kinds of tweets and images are associated with better ratings.

Ratings are political. We find, if we don't remove hashtags and username mentions, these tend to be the most relevant features in the SVM. This aligns with initial observations that the tweets (and ratings) are frequently politically motivated. Indeed, even after stripping out hashtags and usernames, we find "bipartisanship" and "fauci" to be in the top 50 most positively correlated tokens, while "scandal" and "vaccinated" are among the 50 most negatively correlated. Interestingly, in the multinomial naive Bayes model, features associated with high ratings were not hashtags or user mentions for the most part, but those associated with low ratings were, overwhelmingly.

Bare backgrounds are bad. We ran the image classifier on the test set and compared the highest classified (confident good rating) images against the lowest classified (confident bad rating) images. Qualitatively, as seen in Figure 6, we find images which are predicted to receive a bad rating have blank white walls. We compare this to images which are predicted to receive good ratings, which appear to have much more visually interesting backgrounds.

It's the little things. The Naive Bayes model picks up on some specific pieces of decor in highly rated rooms. These may be useful to consider when redecorating: pillow(s), lighting, plant/flowers and art. Other predictor words determined by the Naive Bayes model can be found in Figure 7. As expected, predictor words for below-median ratings tend to have a more negative connotation than those with a higher rating.

8 Evaluation

Evaluation results are included in Table 3. The text classification models all performed roughly equally as well, but multinomial Naive Bayes has the highest performance across all metrics.

In addition, while text classification performed well with basic preprocessing, we attempted to improve evaluation scores through further preprocessing, such as removing stopwords and using bigrams instead of words. Table 4 shows the impact of additional preprocessing on multinomial Naive Bayes and Logistic Regression. While removing stopwords did seem to particularly improve or worsen our models, using bigrams clearly worsened both our Naive Bayes and Logistic Regression models.

Model	F1	AUROC	Accuracy
Text			
Multinomial NB	**0.815**	**0.907**	**0.823**
Linear SVM	0.786	0.871	0.786
Random Forest	0.788	0.887	0.802
Logistic Regression	0.814	0.903	0.818
Image			
ResNet Fine-tune	0.700	0.766	0.669
LDA Topic and Logistic Regression			
10 Topics	0.012	0.495	0.241
5 Topics	**0.035**	**0.509**	**0.295**

Table 3: Evaluation scores for each model tested

The image classification model does not perform as well as the text models, but still picks up on useful signals, which we show both with the evaluation metrics as well as qualitative comparison in Figure 6. However, we do not isolate the background from other elements of the image, such as the interviewee or news ticker. It's possible that these extraneous elements bias the image classification.

The LDA-only logistic regression was less effective compared to our other models. When examining the topics, we also could not discern salient themes; we suspect this may be due to low variance that the room rating tweets have in terms of different topics. They tend to be relatively formulaic and are, of course, all centered around evaluating room decor.

9 Conclusion

9.1 Findings

As seen in Table 3, metrics between different text models are not substantially different; most predict reasonably well (around 0.8 for all metrics). This similarity could be due to politics being a large component of prediction, as well as the small vocabulary size in proportion to the number of tweets (12619 tweets with a vocabulary size of 10,113 words). A narrow pool of words could make it more difficult to

Figure 6: Comparison of highest (left) vs. lowest (right) classified images from test set.

Figure 7: Top 50 predictor words from Naive Bayes for below median ratings (left) and above (right)

Preprocessing	F1	AUROC	Accuracy
Multinomial NB			
Default	**0.815**	**0.907**	0.823
Stopwords Removed	0.811	0.905	**0.828**
Bigrams	0.783	0.881	0.793
Logistic Regression			
Default	**0.814**	**0.903**	**0.818**
Stopwords Removed	0.800	0.903	0.813
Bigrams	0.775	0.876	0.792
LDA: 10 Topics			
Default	0.012	0.495	0.241
Stopwords Removed	0.029	0.506	0.283
Stemmed + Stopwords Removed	**0.039**	**0.512**	**0.299**
LDA: 5 Topics			
Default	**0.035**	0.509	0.295
Stopwords Removed	0.031	0.507	0.291
Stemmed + Stopwords Removed	0.034	**0.512**	**0.302**

Table 4: Neither removing stopwords nor using bigrams clearly improves the models. In fact, using bigrams caused significant decreases in evaluation metrics. However, topic modeling was improved with stemming and removing stopwords.

distinguish between the different classes and lead to more limited models.

The ResNet image classification model also performs admirably at approximately 0.7 for all metrics, but combining LDA Topic Analysis and Logistic Regression did not work well, with an accuracy of only 0.295. This is likely because all of the tweets in the twitter account use similar words and a similar structure. Further exacerbating the similarity of all of the tweets, the @ratemyskyperoom account is either run by one person or a very small group.

Finally, we found that there was not a positive effect between removing stopwords and using bigrams with most of the text classification models and our dataset. It is possible that stopwords are evenly distributed between tweets, due to the similarities already mentioned, which leads to stopwords making little difference. In addition, using bigrams probably had a negative impact, most likely because it simply led to fewer features in an already small vocabulary. With similar and short tweets, bigrams were not effective for text classification. In the case of LDA topic modeling, while further preprocessing did not negatively effect its scores, it did not result in a very high improvement for each score in general, for both 5 and 10 topics.

Overall, we found that there are a couple factors that may help maximize the score @ratemyskyperoom will award your background: (1) be a liberal politician or journalist to match the account's political biases, (2) avoid sitting in front of a blank or boring wall, and (3) keep pieces of decor such as art or plants.

9.2 Further Research

This project raises several open questions. First, rating tweets tend to include both assessment and justification. To derive more useful insights from this data, we'd like to be able to separate the justification from the assessment. For example, we are less interested that tweets with high ratings include the words "beautiful" or "amazing", but more interested in whether they include words such as "art" or "plants".

For our image analysis, further work would include interpreting the model: what exactly is the image classification model looking for? Is it being biased by the subjects in the photos? In line with the political biases present in the text classification, are the news logos providing bias to the image classification model?

Finally, multimodal models which include both the text and image data could significantly improve performance as well as interpretability. Incorporating the topic modeling probabilities into the text based models could also yield better results. It seems clear that the models are picking up on different signals, so this would be a fruitful avenue for further research.

Personal Contributions

Naitian Zhou worked the image and text classification models. He has an ugly room: bare walls, sterile lighting. 2/10.

Parth Dhyani worked on the data exploration and topic modeling. He sits in an amusing room with a blue ceiling painted with clouds that would fit better in a toddler's playroom. 6/10.

Wen Si worked on data preprocessing & exploration and some text classification models. Her room has a mirror door that always shows her and her roommate's unmade beds. 2/10.

Drishaan Jain worked on methods background research and data exploration for popular words and engagement. His room features vintage travel posters and a wall of photos of his friends. 8/10.

Isha Hameed worked on LDA topic modeling and some background research. Sadly, her dimly lit green room features only 1 framed picture of a bird. 3/10.

Acknowledgments

We acknowledge you.

References

J. Gautam A. Goel and S. Kumar. 2016. Real time sentiment analysis of tweets using naive bayes. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 257–261, Dehradun, India.

J. H. Bappy, J. R. Barr, N. Srinivasan, and A. K. Roy-Chowdhury. 2017. Real estate image classification. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 373–381.

Liam Bullingham and Ana C. Vasconcelos. 2013. 'the presentation of self in the online world': Goffman and the study of online identities. *Journal of Information Science*, 39(1):101–112, Feb.

Jose Eurico de Vasconcelos Filho, Kori M. Inkpen, and Mary Czerwinski. 2009. Image, appearance and vanity in the use of media spaces and video conference systems. In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, page 253–262. Association for Computing Machinery, May.

Erving Goffman et al. 1978. *The presentation of self in everyday life*. Harmondsworth London.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Matthew K. Miller, Regan L. Mandryk, Max V. Birk, Ansgar E. Depping, and Tushita Patel. 2017. Through the looking glass: The effects of feedback on self-awareness and conversational behaviour during video chat. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 5271–5283. Association for Computing Machinery, May.

Yihong Zhang and Adam Jatowt. 2019. Image tweet popularity prediction with convolutional neural network. In *European Conference on Information Retrieval*, pages 803–809. Springer.