Customer Satisfaction by Classifying Emotion

Noah Hendrickson, Eric Trempe

1. Introduction

Section 1.1 Motivation

In the world of business, customer satisfaction is everything. Bad reviews or unhappy customers can cause even the strongest company to crumble. It is incredibly important that companies take customer feedback they are given and improve from it. “Customer satisfaction is one of the few levers brands can still pull to differentiate themselves in crowded and competitive marketplaces. Today, the brand with the best customer experience usually wins.”[2]

There are many ways to gauge customer satisfaction. A common way is to provide customer satisfaction surveys. Customer satisfaction surveys, also known as CSATs, are used to determine a score representing how satisfied a customer was with an experience. Various questions are asked about the experience, and the customer is asked to rate their overall experience. This is often measured on a 1 to 5 scale, with 1 being least satisfied and 5 being most satisfied.[3] The problem with this method arises when customers cannot, or choose not to, take the survey. Often, only 10-30% of people fill out those satisfaction surveys, which means valuable information to the company is walking out the door.[4] Additionally, there are issues with voluntary response bias and inconvenience that hinder both customers and employees. In order for a company to properly gauge customer satisfaction in order to make changes, they need to be able to determine how satisfied those customers were with their experience. We propose a method to determine customer satisfaction, and, by proxy, fill out those customer satisfaction surveys, based solely on video that can easily be obtained through security cameras or any sort of video recording devices that are common in stores by classifying the emotion that the customer is experiencing. While obviously customer consent is required for this method as well, they do not need to spend the time to fill out the survey and there is no time between their original reactions and the capturing of emotion. Since this machine learning process is automated, it takes care of the majority of voluntary response bias and other human related inaccuracies.

Section 1.2 Methods

It is widely studied that emotion has a huge effect on how a customer interacts with a company and its products. Customers that are more emotionally connected buy more products, visit more often, and are “twice as valuable as highly satisfied customers.”[1] Customers that have bad experiences will not be as satisfied and may not come back at all. We decided to use emotion as a satisfaction estimator because it is relatively simple to detect and usually is a good indicator of how someone is feeling in the moment, which we believe maps well to satisfaction. The only problem with using emotion in this way is that there is no 1-1 mapping of emotion to a customer survey. Because of this, we developed our own method in order to fill out these surveys with the detected emotion. There are 7 emotions we base our output off of: happiness, neutrality, surprise, sadness, disgust, fear, and anger. Using these emotions, we took the AFINN[8] and NRC[9] lexicon sentiment scores for these words, which ranged between -3 and 3, and converted them

proportionally into a 1-5 scale. We then use the probabilities for each emotion output by the classification model (section 2) to weight and sum the scores into a final satisfaction rating. One potential problem we anticipated with this method is the lack of a baseline emotion. Some customers could have outside factors influencing their feelings at the time of the survey. Our solution to this problem was to model the satisfaction score over time with different weights. When going through an experience, one of the most influential points of the experience when it comes to how you feel about it is the end–one half of the peak-end rule.[11] Because of this, we gave larger weights to the frames towards the end of the experience. We used a slightly modified softmax on the frame count which was then applied to the scores for each frame to get the scaled weights. A tensor is generated containing numbers from 1 to N, N being the number of frames in the video. Next, softmax, with the input scaled by 0.03, is applied to the tensor and the resulting weight tensor is applied as a dot product with the frame score tensor to produce a score for the video as a whole. This method allows for the last 30 or so frames of the input video to still meaningfully contribute to the overall satisfaction score. The final modification to the output score is the second half of the peak-end rule: the peaks.[11] The max peak and the min peak values are found over the frames and weighted based on which comes first. The peak that comes earliest is weighted by 1/10. The peak that comes latest is weighted at 4/10. The score that was run through the softmax weighting is weighted at 5/10. The output score is the sum of each of those 3 weighted scores.

2. Architecture

The determining emotion part of our problem boils down to a machine learning classification problem. We need to first identify frames with faces from a video, then classify them as a combination of emotions. Because we have image data, convolutional neural networks seemed to be the best method to test.

We tested three models total. Each model was convolution based and the third model had built in facial detection for increased classification performance. The first model we tried was a basic network based on the LeNet architecture with 6 convolution layers, 4 Max Pools, ReLU activations connecting the convolutions, and 3 linear layers going into the output.

The second model we tried was a model based on the GoogLeNet architecture presented in the paper "*Going Deeper With Convolutions.*"[7] This model made use of "inception blocks" which consist of side by side convolutions. In a single inception block, the input is put into four side-by-side convolutions. The first is a 1x1 convolution into a ReLU into a batch norm. The second is a 3x3 into a batch norm/activation into another 3x3 convolution into a batch norm/activation. The third is the same as the previous but with 5x5 kernels. The last is a max pool into a 1x1 convolution into a batch norm/activation. Once the input has been run through each of these individually, each result is concatenated together and passed on. The main net's architecture consists of an initial 7x7 convolution then a 3x3 convolution, both followed by Local Response Norms and activations. Then follows a single inception block into a dimension reducing max pool into a second inception block again followed by a dimension reducing max

pool followed again by a third inception block. The output of that is put into an average pool in order to flatten to a 1xn vector. Following this is a dropout layer with probability of 40% which is key to reduce overfitting. The model finishes with a linear layer and final output softmax.

The final model that we tested was a combined face detection / emotion prediction model. The emotion prediction portion of the model utilizes a technique called "Residual Masking", described in the paper *Facial Expression Recognition Using Residual Masking Networks*. This is done in residual masking blocks which have two parts: a residual part, and a masking part. The residual part produces a coarse feature map and the masking part produces an activation map. Each block is designed in a U-net structure with one encoding path and one decoding path.[6] The backbone of the residual portion of the model is Resnet34.[10]

All of these models return a vector of 7 probabilities, one for each emotion, which is then transformed using the weighted sum from section 1 to give our satisfaction score.

3. Training

Both our rendition of GoogleNet and LeNet were trained from scratch on the FER2013 dataset. This dataset contains 28,709 training samples and 7,178 testing examples consisting of 7 emotions: happy, sad, angry, neutral, surprise, disgust, and fear. Some categories of emotions were underrepresented, so in order to curb this we did some data augmentation during training, including a 50% chance for a horizontal flip and a random crop of the image down to 40x40 pixels. Both models were trained until either convergence– defined by the train set accuracy not improving for 10 consecutive epochs–or when the maximum 50 epochs were reached. After hyperparameter tuning, both models were trained using the Adam optimizer and a learning rate of 0.001. These parameters were more successful in terms of validation accuracy than stochastic gradient descent with momentum and learning rates of 0.0001 and 0.01. A batch size of 50 was used as it was a good balance of faster training and GPU memory use.

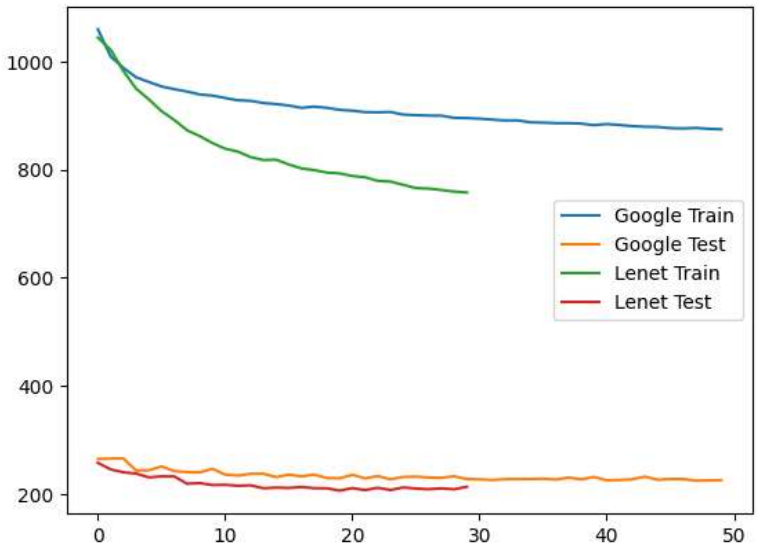
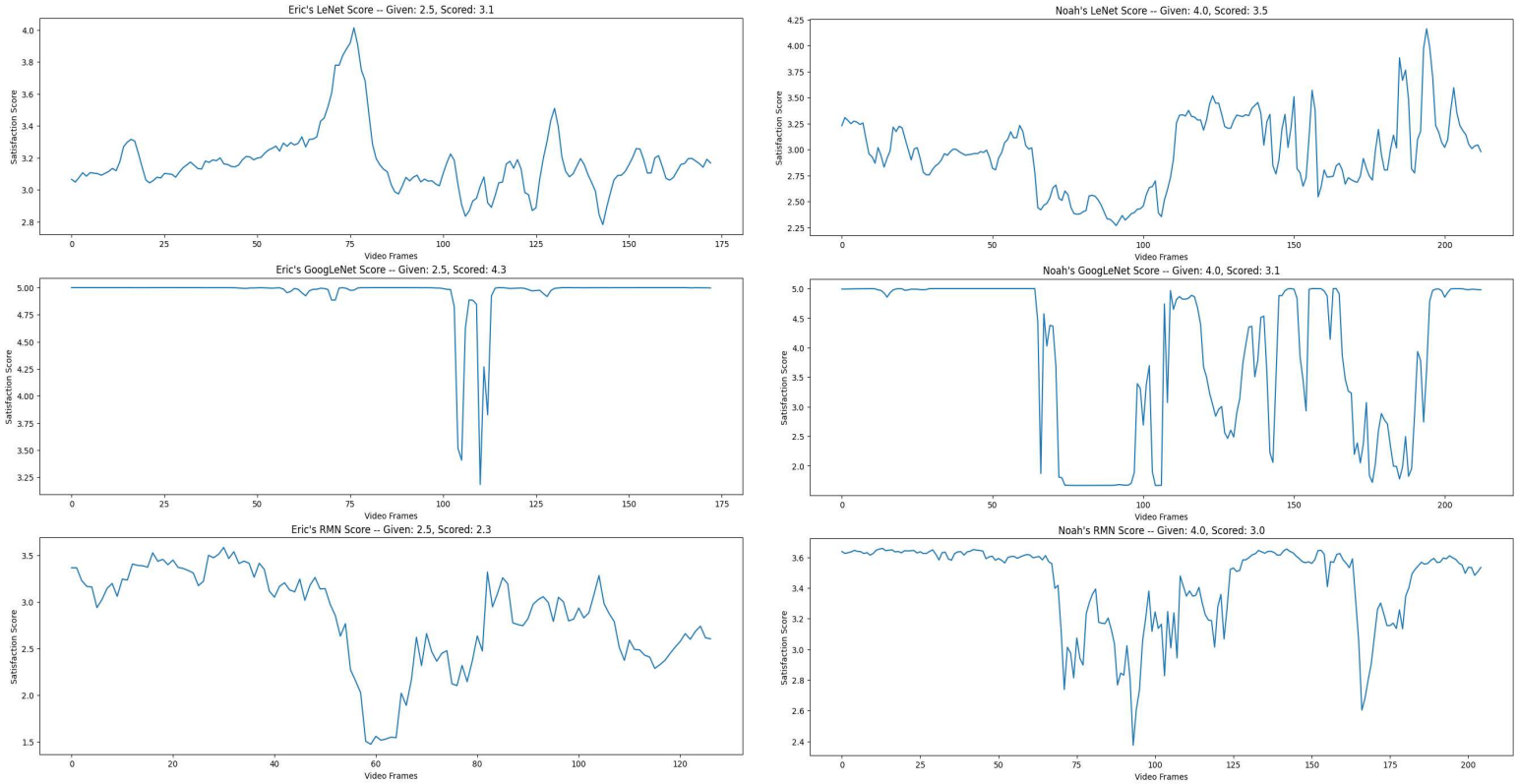


Figure 1: Graph of the training loss for our GoogleNet and LeNet renditions over the epochs they were trained. Our rendition of LeNet converged before the 50 epoch max.

4. Results

Testing our method proved to be difficult because this idea has not really been explored before. There are not really any available videos of people with corresponding satisfaction scores, so in order to test our method, we gathered videos of people expressing emotions and their hypothetical satisfaction rating as if that scenario were to occur in a consumer setting. Below are two examples of the model classifying emotion into satisfaction ratings over the course of a video.



We tested 8 videos total[1]. All of the videos and scores can be seen in this google drive folder where the video file names are structured as *[name][given score].[extension]* and the 'p' characters in the titles denote a decimal point. Results from each video can be found in the *[name]_results.png* files. Based on the validation classification accuracies, 44%, 59%, and 74% for our LeNet, GoogLeNet, and RMN models respectively, we expected the RMN model to most closely match the user emotion ratings and our examples show that that is the case. For these 8 examples, the LeNet model was off by an average of 1.1, GoogLeNet was off by an average of 1.775, and RMN was off by an average of 0.7625. The three graphs for each video follow roughly the same peaks but there is much more variation in the rest of the videos. It is also interesting to us that GoogLeNet has a much worse average difference than LeNet. This suggests to us that perhaps as the model classifies emotion more accurately, the need for a face detection

[1] All the linked videos are of group members or people members of our group know, and the videos were given with explicit permission to be shared in this report.

model leading into it grows. The GoogLeNet model could be picking up on parts of the images that are not faces, affecting the accuracy, while the LeNet is focused on the proper aspects of the image. This also supports the RMN model performing the best since it has a face detection model built in, and therefore would not get distracted by other parts of an image.

5. Conclusion

In conclusion, our method of determining satisfaction scores from videos obtained of customers is not perfect. The variability of natural facial expressions of customers is likely to impact the results of certain individuals. For example, with the RMN, in Noah's video, it can be seen that they express satisfaction, but their facial expressions may suggest to the model that they are expressing a more negative emotion. On the other hand, Eric's video shows more traditional emotion expressions and the score almost exactly matches the given score. That is one of the major downsides of this way of determining satisfaction. Everyone expresses emotions in different ways and our method may not capture that difference for everyone, which was also pointed out in our feedback reports. Training on images containing a wide variety of human faces is the best attempt we can make at removing biases related to this problem. Finally, the models only classify 7 emotions, but humans can express many many more emotions in reality.

Section 5.1 Future Work

In the future, in order to perhaps make our method more robust, we could find a dataset with more emotions and expand our emotion-to-value mapping to make it more fine grained and capable of realizing more in-depth and accurate scores. More emotions and additional faces to train on would always be helpful in making a model that captures more of the nuance in human emotions. With this addition, we believe our model could be a quick and accurate way to supplement real satisfaction surveys. We could also look into deeper networks given more time and resources to attempt to increase emotion detection accuracy. Finally, adding our own face detection model could improve the accuracy of our emotion detection models as well.

Section 5.2 Incorporated Feedback

Much of the feedback we received was related to using a pre-trained model for facial recognition. We did end up deciding to use a pre-trained model, the RMN, that we compared our models for this purpose. But, in future work, piping a face detection model to our other models may be beneficial. Other feedback was centered on the privacy concerns with this type of video feedback. While that is a concern, there is already volunteerism required to get customer feedback with more traditional methods, like surveys, so we believe the convenience and practicality of this idea will encourage participation. Finally, we received many questions about how this could replace survey scores, which we considered when developing our weighted average models. We made sure to use lexicons and weightings with academic support to ensure that we interpreted emotions in an accurate manner to get accurate consumer feedback.

References

[1]https://hbr.org/2016/08/an-emotional-connection-matters-more-than-customer-satisfaction (customer satisfaction survey info)

[2]https://www.helpscout.com/blog/customer-survey/ (customer satisfaction survey info)

[3]https://www.qualtrics.com/experience-management/customer/satisfaction-surveys/ (customer satisfaction survey info)

[4]https://peoplepulse.com/resources/useful-articles/survey-response-rates/#:~:text=Customer%20satisfaction%20surveys%20and%20market,respondents%20who%20complete%20your%20survey.” (customer satisfaction survey info)

[5]https://github.com/phamquiluan/ResidualMaskingNetwork (Residual Masking Network)

[6]https://ieeexplore.ieee.org/document/9411919 (Residual Masking Network Paper)

[7]https://arxiv.org/abs/1409.4842 (GoogLeNet paper)

[8]http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010 (AFINN lexicon)

[9] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (NRC lexicon)

[10]https://pubs.aip.org/aip/acp/article/2484/1/060010/2879736/Comparison-of-different-convolutional-neural (Resnet34 paper)

[11]https://positivepsychology.com/what-is-peak-end-theory/ (peak end rule)