# Zero-Shot Emotion Classification for Determining Customer Satisfaction

Noah Hendrickson

May 1, 2023

## Problem Statement

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.

## Methodology

In a previous project[8], we proposed a method to determine a satisfaction score for a video that utilized convolution networks, specifically a GoogLeNet variation like that presented in *Going Deeper With Convolutions*[5], a Residual Masking Network, presented in *Facial Expression Recognition Using Residual Masking Network*[6], and a LeNet rendition like that presented in *Gradient Based Learning Applied to Document Recognition*[7]. Seven emotions–happy, sad, angry, neutral, disgust, surprise, and fear–were classified using these networks, and the probabilities obtained from each network were then used to weight scores for each emotion in order to determine a score for an image. 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] Satisfaction score for a whole video was enabled through utilizing the peak-end rule which states that, when looking back on an experience, how you feel about it is most impacted by the peaks and the end. In order to integrate this, a score was calculated for each frame in the video. A slightly modified soft-max was then applied to a tensor of the frame numbers in order to generate weights for each frame of the video. This allowed for the last 30-50 scores in the video to apply most satisfying the end portion of the peak-end rule. In order to consider the peaks, the max and min peaks were found and a weighted average of the two peaks and the total score found from the soft-max weight dot product of the frame scores was taken to get the final output score.

This process, however, came with a few limitations. The emotions in the FER2013 dataset were not fine grained enough to really get a good idea of how multiple emotions on a face at once may affect the score the person would give, which led to inconsistent and jumpy results. Additionally, the convolution networks had to be trained on a gray-scale dataset with only about 7,000 images. This led to not great classification accuracy.

In this project, I will attempt to determine satisfaction score through transfer learning by utilizing OpenAI's pretrained CLIP Processor and Model to get image embeddings of the images I wish to classify and text embeddings of the possible emotions that I will want to consider and find the cosine similarity between the two in order to determine which emotions the image is expressing. This will take place of the convolution networks in the previous problem. This will allow me to arbitrarily specify the emotions that I want to consider, and also be more specific about what sort of emotions. The CLIP model also produces probailities of each of the text embeddings, allowing the same classification-to-score process to happen.

## Datasets and Result Visualization

For this problem, there isn't really a dataset that I can use to validate the method. For the emotion classification portion, I can use the FER2013 dataset in order to validate that finding the cosine similarity between the embeddings works as a classification method. For this purpose I also will use a dataset I found on Kaggle called "Natural Human Face Images for Emotion Recognition."[9] For the actual scoring portion I can reuse videos that I collected to test on the score generated by the CLIP model. Graphs and visualizations in the report will be mainly comprised of accuracy on the FER2013 dataset of the CLIP model and the three models tried in the previous project, as well as graphs of the score over time of a video and the final scores of each video with comparisons to the given scores.

## References


[1] An Emotional Connection Matters More than Customer Satisfaction

[2] Customer Satisfaction Surveys: A Comprehensive Guide

[3] Customer satisfaction (CSAT) surveys: Questions  template

[4] SURVEY RESPONSE RATES: Tips On How To Increase Your Survey Response Rates

[5] Going Deeper with Convolutions

[6] Facial Expression Recognition Using Residual Masking Network

[7] Gradient Based Learning Applied to Document Recognition

[8] Customer Satisfaction by Classifying Emotion (Google Doc)

[9] Natural Human Face Images for Emotion Recognition