Zero-Shot Emotion Classification via CLIP for Customer Satisfaction Scoring

Noah Hendrickson

May 5, 2023

1 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. In order to solve this problem, I classify emotion from videos and use the emotion probabilities to determine a score for the video.

2 Methods and Resources

2.1 Emotion Classification

There are two parts to this problem, the emotion classification part, and the scoring of videos. For the first, emotion classification, I utilized OpenAI's CLIP model, along with previously tested convolution models based on GoogLeNet, ResNet, and LeNet to test against.[5][6][7] I will not describe in detail the convolution models, as they are just to benchmark the CLIP model, but more details can be found in the other report preceeding this one.[8]

OpenAI's CLIP is a pretrained model, oft used for zero-shot classification tasks or text-to-image generation tasks. The model has two parts: a processor that takes text labels and images and converts them to embeddings and attention heads, and a model portion that takes those embeddings and calculates similarities. The output values that are relevant to this problem are the logits that determine which text labels are represented in the image. Those logits can then be softmaxed to get label probabilities.

The key reason that I decided to use CLIP for this problem is the ability to compare both images and text, as well as the fine grained control over the emotions able to be classified. With the convolution models, only 7 emotions were able to be represented, those from the datasets described in section 2.3, leading to not very smooth results. With CLIP, you can specify as many emotions as you want, which means more control over what is being seen by the model and more depth into what things the customer might be feeling in the video.

2.2 Frame/Video Scoring

The second part to the problem, scoring of the videos, is done via a custom process that doesn't involve any models besides inferencing from CLIP. First, labels are created for which emotions want to be classified in the video. For the purposes of this project, I utilized the following emotions under broader emotion categories:

- Angry: angry, annoyed, mad, enraged
- Happy: happy, delighted, joyful, content, exhilarated, bliss, delighted, gleeful
- Sad: sad, miserable, dejected, down
- Surprised: surprised, shock, astonished, horrified, astounded, stupefied
- Disgusted: disgusted, revolted, offended
- Neutral: neutral, straight-faced, disinterested, intrigued, impartial, unbiased, open-minded
- Fear: fear, scared, terrified, off-put, horrified, afraid, anxious

Each emotion is assigned a score between 1 and 5. A score of 1 is generally represented as negative emotions such as angry, enraged, disgusted, or horrified. A score of 5 is generally represented as positive emotions such as delighted, bliss, or exhilarated. Then a video is taken of a person expressing some emotion that is to be captured for the score. The video is read in frame by frame, each frame being embedded alongside the given emotion labels using the CLIP Processor. Embeddings and attention heads are then passed through the CLIP Model to get the logits output which are then softmaxed to obtain probabilities of each emotion. The vector of probabilities for each emotion and the vector of scores for each emotion are then dot-product'd together in order to obtain a single frame score.

The final part of the process is to convert the scores of each frame in the video to a single score. That is done through a custom process that attempts to incorporate some information about how humans process experiences. The peak end rule states that, when looking back on an experience, people most remember the end and the peaks of the experience. The "end" half of this rule is incorporated into the scoring through use of a modified softmax function. First, all of the frame scores are obtained and stored. Another vector is then created that contains the frame numbers in ascending order. A modified softmax function is applied to that vector to obtain weights for each of the frame scores in the video. Like with the probabilities, this vector and the vector of frame scores are then dot-product'd together to obtain an "end-score." The softmax is modified in such a way that roughly the last 30-50 frames will contribute the most to the final score. The "peak" part of the peak-end rule is convered by finding the maximum and minimum scores in the video. The final score is obtained by doing a weighted average of those two "peak-scores" and the "end-score." The weight for the end score is always the same at $\frac{5}{10}$ and the weights for the two peaks are $\frac{1}{10}$ and $\frac{4}{10}$ depending on which came earlier and which came later respectively.

2.3 Datasets

For this problem, two datasets and a small set of custom videos were tested on. The first, FER2013, is a dataset of roughly 7000 48x48 pixel, grayscale, images depicting faces of people expressing emotions and their labelled emotion.[9] The emotions for this dataset are angry, happy, sad, disgust, surprise, neutral, and fear. The second, Natural Human Face Images for Emotion Recognition, is much like FER2013 in that there are roughly 5000 images of grayscale faces depicting the same 7 emotions, however, the images are larger, around 200x200 pixels.[10] This importance will be discussed more in the results section. This dataset also has an 8th emotion, contempt, however, I decided to not include this to keep this dataset consistent with the labels of FER2013. The final dataset is a set of 8 videos obtained from myself, group members from the previous report, and people we know. Each video varies in length and emotion represented and each comes with a given score that each individual rated the "experience" as. Each video was included in this and the past report with express permission of the individuals included in them.

3 Results

3.1 FER2013 and Natural Human Faces Datasets

First, in order to gauge how the prompts for the CLIP labels should be structured in order to get the best accuracy, I tested against the FER2013 and Natural Human Faces datasets. The first I tested on was FER2013. At first, I had just the 7 emotion names themselves as the labels. The resulting accuracy of this trial was better than random guessing at roughly 13%, however, not anywhere close to the accuracy of any of the convolution models. For reference, on FER2013, the LeNet model got roughly 44%, the GoogLeNet go roughly 60%, and the ResNet got roughly 75%. In order to improve this, I tailored the prompt to better fit the dataset, going from just the emotion name to "gray person expressing {emotion}." This dramatically improved the accuracy rate, going from the roughly 13% to 42%. This is not quite as accurate as the ResNet, however, high accuracy isn't a super important metric in regards to this problem.

I then tested on the Natural Human Faces dataset. For the sake of time, I only tested the ResNet and the CLIP Model. For this dataset, using just the 7 emotion labels, the CLIP model got 46% accuracy and the ResNet got 48% accuracy. This is a dramatic shrinking in difference between classification accuracy, showing that CLIP does much better on larger images with more detail. This makes sense as there's more information to extract to compare against the label embeddings. However, this was not the end of my testing on this dataset. In order to determine that adding more fine grained emotions would be a viable strategy, I classified on the same dataset, but expanded the emotions to contain all mentioned in Section 2.2. In order to still do classification, I tried two separate methods: summing the probabilities of each sub emotion to get the total emotion probability then doing the normal classification, and classifying the image based on the emotion that the sub-emotion with the highest probability falls under. The confusion matrices for the two are represented in **Figure 1** and **Figure 2** respectively. For the summing-probabilities classification method, the resulting accuracy across the dataset was 43% and for the max-probability classification the resulting accuracy was 40%. Both of these accuracies are very close to the original 46% and are especially impressive considering the raw number of emotions that are being classified here. Looking at the summation method confusion matrix in **Figure 1**, we can see that the model was not very good at classifying disgust or sadness properly, but excelled at surprise, neutrality, and, to a lesser degree, happiness and anger. Looking at the summation method confusion matrix in **Figure 2**, we can get a better idea of which specific emotions are being correlated to the overall category. It can be seen that happiness was often correlated with neutral emotions and surprise was often correlated with fearful emotions, but both of those conclusions make sense as its very easy to see how the facial expressions of those two emotions match. Overall, the performance on these datasets showed to me that this is a viable option to try to use on real videos.

3.2 Video Scoring

As mentioned in Section 2.3, 8 videos were collected to test on. Each video was put through each of the 4 models–LeNet, GoogLeNet, ResNet, CLIP–and a score was calculated. For the sake of not making this a 10 page pdf, the result graphs are included in a google drive folder that can be found at this link here as well as the videos corresponding to each graph. The graphs consist of the frame number on the x-axis, and the score for the frame on the y-axis. The reported and classified score are given at the top next to the name which corresponds to the video. Looking at the graphs and the videos, we can see that, in general, the CLIP model more accurately follows the peaks of the videos, in addition to having a much smoother and far less jumpy line, both of which I think can be seen best in Eric and I's results. This is not the case for *every* video, however, because of the imprecise nature of emotions, I am not expecting that it would. We can see from **Table 1** that the mean average error of the CLIP model is better than any of the other three models. The one that gets closest is actually the LeNet model which was by far the smallest and lowest accuracy of the three convolution models.

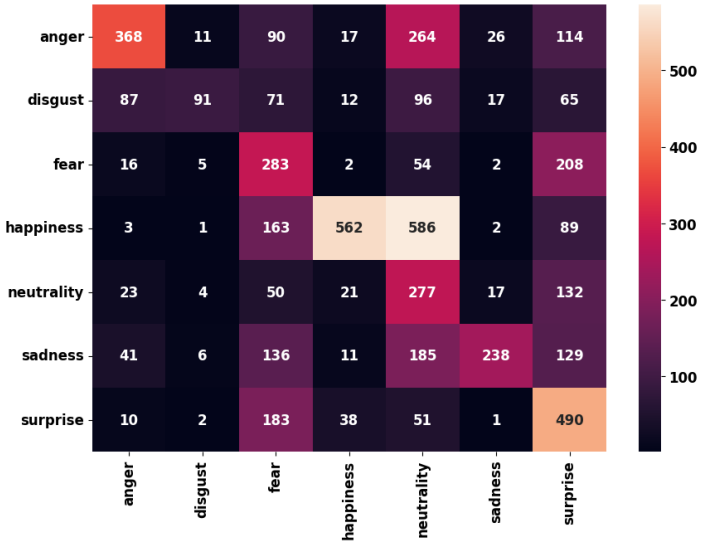

Figure 1: Confusion matrix for testing on Natural Human Faces dataset with summing probabilities classification method.

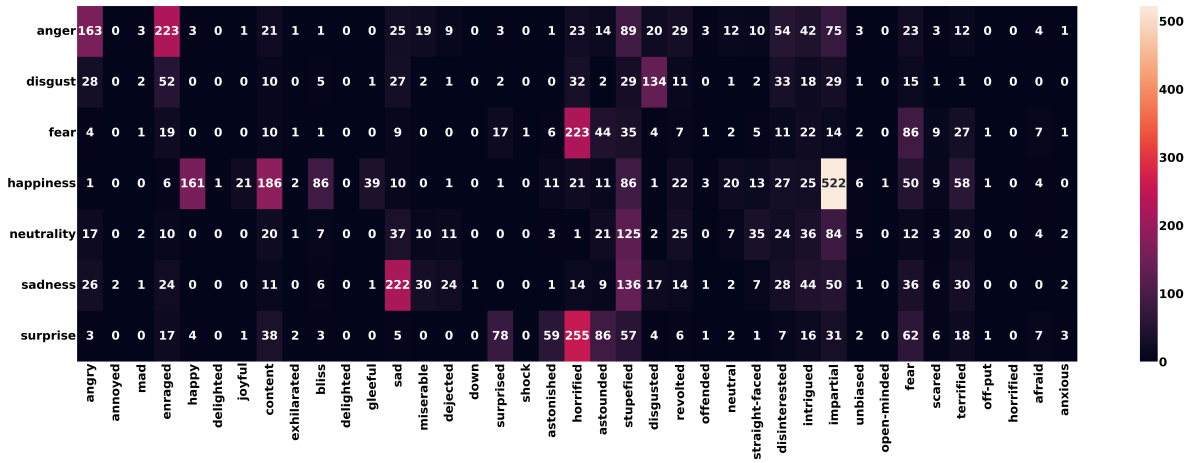

Figure 2: Confusion matrix for testing on Natural Human Faces dataset with max probability classification method.

Model	MAE
LeNet	1.1
GoogLeNet	1.74
ResNet	1.4
CLIP	0.96

Table 1: Models tested on the 8 videos along with their mean average error scores.

3.3 Discussion

These results are far better than I expected to get. The CLIP model by far performs the best of the models while also taking into considering far more emotions than any of the others. The fine grained control over the emotions also allows the process to be tuned to more fit the setting. If its a food testing setting, the person running the model may want to include mroe food related emotions such as disgust, satisfaction, delight, or other ones. This is a powerful tool for the model to have for this problem. Additionally, the model does not have to retrain each time a new emotion needs to be added. Its as simple as adding the emotion to the dictionary of emotions and passing it through.

Another interesting result is that of the LeNet vs. the CLIP model. The two's MAE scores are very close with only a difference of 0.14 between the two. This suggests to me two possible scenarios. The first is that the LeNet model may have just gotten lucky on these videos. Eight videos is not a very large sample set, and its very possible that with a larger dataset of videos, the LeNet model could start to fall behind much more. The second thing that it suggests to me is that a high accuracy is actually more of a detriment to the scoring, rather than a benefit. Its very possible that, because of the incredibly range of emotions that humans can express at one time, training a model to just recognize a single emotion in a face will never be a good idea for this problem. I hypothesize that its much better to have a lower accuracy model that better captures the many different emotions that can be showing on a face at once, to provide a far more detailed and accurate satisfaction score overall.

4 Ethics Considerations

This problem comes with inherent ethical considerations. Most of the ethical considerations are in relation to the actual problem itself, rather than the models used in the problem, though there is a small consideration there that I will start with. When it comes to the CLIP model, we are unsure what material it was trained on, some of which could be images of people who may not have consented. With the convolution models, they were all trained on images of people from datasets. These datasets could also have images of people who did not consent to be in the dataset. These datasets may also be unbalanced in terms of representation. I was not able to check the datasets for population representations, however, there's a non-zero chance that many groups, such as women or POC, may be underrepresented in these datasets, leading to worse performance on videos of these individuals.

The main ethical consideration for this problem is the gathering of videos in the first place. In many places in the US, consent is required to record people, but how does that apply to a company? Are companies able to take security footage and do whatever they want with it? Is it morally right to attempt to classify a satisfaction score from a video of a person? I have no concrete answer, but I do have arguments on each side. On one side, this method couldn't be used to harm really anyone. Nothing is being published about the emotions of the person nor is anything probably even leaving the hypothetical company that this might be used at. All that the model is doing is classifying and getting a satisfaction score. On the other side, it is a breach of privacy. Nobody wants videos being taken of them without their consent and used to influence things, even if it is just a score gotten from a video. Trying to guess at someone's opinions also seems a little weird? Yes, nothing is being done with those videos, but what if some data is leaked? Then there are videos of you out there, and possibly your alleged emotions and opinions classified my a ML model. There is no easy answer to this issue, but it is an important thing to consider.

5 Conclusion

In conslusion, this project not only shows that this method of determining customer satisfaction is viable, but it also shows the power of zero-shot learning and its application. In the future, I'd like to explore more intricate ways of calculating the satisfaction score in order to refine the process even more, as well as collect more videos to test on to get an even more accurate view of how well this method works. For all the model and testing code, you can follow this link and, again, for the videos and results for each video, you can follow this link. Both are also linked in the references. Thanks for the wonderful semester, I had a ton of fun and hope to take some of the things I learned to my future classes and work!!

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] FER2013

[10] Natural Human Face Images for Emotion Recognition

[11] Peak-End Rule

[Code] Model/Testing Code

[Videos] Videos and Video Results