Black-Box Emotion Detection: On the Variability and Predictive Accuracy of Automated Emotion Detection Algorithms

Francesc Busquet
University of St Gallen
Christian Hildebrand
University of St. Gallen

Cite as:
Busquet Francesc, Hildebrand Christian (2020), Black-Box Emotion Detection: On the Variability and Predictive Accuracy of Automated Emotion Detection Algorithms . *Proceedings of the European Marketing Academy*, 49th, (64028)

Paper from the 49th Annual EMAC Conference, Budapest, May 26-29, 2020 .

Black-Box Emotion Detection: On the Variability and Predictive Accuracy of Automated Emotion Detection Algorithms

Abstract

The ubiquitous availability of image data, advances in cloud-computing, and recent developments in classification algorithms gave rise to a new class of automated emotion detection systems which claim to perform accurate emotion detection from faces at scale. In this research, we provide a tightly controlled validation study using pretrained emotion detection algorithms of the Google ML, Microsoft Cognitive Service, GfK EmoScan, and other platforms to test the robustness and consistency across and within current emotion detection systems. Our results demonstrate considerable variability in predictive validity across emotion detection systems, high variability across different types of discrete emotions with strong positive emotions (such as an open mouth smile) being easier to classify compared to negative emotions such as anger or fear, and we detect sizable positive correlations of theoretically opposite emotions (such as surprise and fear). We provide two modeling strategies to improve prediction accuracy by either combining feature sets or by averaging across emotion detection systems using ensemble methods.

Keywords: Emotion detection, discrete emotions, machine learning.

Track: METHODS, MODELLING & MARKETING ANALYTICS

Black-Box Emotion Detection: On the Variability and Predictive Accuracy of Automated Emotion Detection Algorithms

Extended Abstract

1. Introduction

Automated emotion detection from facial expressions refers to the use of algorithms that detect facial landmarks in pictures to classify people's discrete emotions (Liu et al. 2014). These automated emotion detection systems classify discrete positive emotions such as happiness or surprise to negative emotions such as anger and fear. A computer vision algorithm first identifies facial landmarks in a picture and then assigns each picture a discrete emotion label based on the features or composition of the existing facial landmarks. Companies such as Microsoft, Google, or GfK provide platforms to perform such automated emotion detection with recent industry reports suggesting a CAGR of 32.7% and a market size of 25 billion by 2020, highlighting the importance and dominance of these AI-powered technologies for the future of marketing.

Despite the increasing availability of automated emotion detection systems, fundamental methodological questions arise. Are automated emotion detection systems valid? Is the same picture classified correctly across emotion detection systems? These questions are important as emotion detection systems are using pre-trained algorithms to classify discrete emotions from facial expressions in a way that is unknown to the user of these systems. Thus, it is likely that the same picture is assigned a different discrete emotion conditional on the type of emotion detection system being used. To the best of our knowledge, both a formal test of this hypothesis is non-existent as is a formal analysis that unravels under which conditions emotion detection algorithms would increase in predictive accuracy.

2. Theoretical Background

Emotion detection from facial expression has been studied predominantly in computer vision and pattern recognition. One of the first lines of research was concerned with the high dimensionality of facial images, or more specifically the facial features in images. Thus, the majority of prior work was focused on dimensionality reduction techniques beginning with distances and ratios among feature points (Kanade 1973) and moving towards more complex methods such as autoencoder networks (Cottrell and Metcalfe 1991) or principal components (Padgett and Cottrell 1997) or, more recently, signal processing methods such as Gabon filters (Bashyal and Venayagamoorthy 2008).

Yu and Zhang (Yu and Zhang 2015) provide evidence that automated emotion detection models reach a 61.29% accuracy on the test set using a multiple deep network learning, while Levi and Hassner (Levi and Hassner 2015) achieved a 54.56% by using Convolutional Neural Networks and so called mapped binary patterns. Even though the accuracy seems similar, results are hardly comparable across studies due to differences in stimuli (i.e., pictures) and modeling techniques (deep neural networks vs. binary patterns). Taken together, due to varying stimuli and methods being used, the current research on automated emotion detection systems provides mixed evidence on the effectiveness across discrete emotions and the dependence on the modeling strategy used. The current paper fills this gap by providing a tightly controlled validation study comparing the effectiveness (i.e., in terms of predictive accuracy) of a variety of automated emotion detection systems across discrete emotions, dependencies on the modeling technique being used, the stimuli set, and ways to improve predictive accuracy by combining feature sets or using ensemble methods.

3. Dataset

We used a standardized picture set with objective ground truth, i.e. knowledge about the discrete emotion that is displayed by actors in a standardized picture set. Specifically, we used the Chicago face database (Ma, Correll, and Wittenbrink 2015) to evaluate automated emotion detection systems. The Chicago face database contains photos from 597 male and female targets of varying ethnicity between 18 and 40 years under standardized conditions. For a subset of 158 targets, images display either a neutral, angry, fearful, or two positive emotional states (happy face with either an open or closed mouth).

Each of these images of the 158 subjects displaying various emotions was classified using a variety of emotion detection Application Programming Interfaces (API): Microsoft Cognitive Services, Google ML, Sightcorp, Kairos, and the GfK EmoScan. The data was normalized across emotion detection systems to provide meaningful comparisons across APIs.

4. Baseline: Accuracy of Single Predictor Models

To set a baseline, we evaluated the effectiveness of the different APIs by taking only a single variable to predict emotion labels. We trained a multinomial logit model to predict each discrete emotion across APIs. Prediction accuracy varied moderately from 49% to 58% with Google's API performing best, achieving a test accuracy of 58.94% and a test Kappa of 40.25% (mean CV accuracy of 58.04% and mean CV Kappa of 39.17%). On the contrary, GfK EmoScan realized the lowest predictive accuracy with a test sample accuracy of 49.81% and a test Kappa of 36.88% (mean CV accuracy of 48.25% and mean CV Kappa of 34.75%). This difference stems primarily mainly from a greater predictive power of the neutral class by the predictions done using Google ML (80.34% balanced test accuracy on neutral class, while the GFK EmoScan model presents a 70.19% balanced test accuracy for that class).

Positive emotions were predicted consistently better across APIs while we observed substantial variation across negative emotions (standard deviation of the balanced accuracies across emotion detection systems: $\sigma_{\text{happy open}} = 3.81$, $\sigma_{\text{neutral}} = 6.89$, $\sigma_{\text{fear}} = 14.98$, $\sigma_{\text{anger}} = 30.9$). Likewise, we observe substantial variation in intraclass correlations (correlations of emotion detection systems within each discrete emotion, i.e. correlation of happiness measures, anger measures etc.) with $r = 0.82$ for happiness, while only $r = -0.239$ for sadness, $r = -0.08$ for surprise, $r = -0.176$ for anger and $r = -0.109$ for disgust, respectively. Moreover, the predictive power of the different APIs for negative emotions was significantly lower than for positive and neutral emotions (average balanced accuracy for the different emotion classes: $\mu_{\text{happy open}} = 78.97$, $\mu_{\text{neutral}} = 79.97$, $\mu_{\text{fear}} = 65.45$, $\mu_{\text{anger}} = 44.3$).

5. Within API Accuracy

To expand the previous results, a subsequent analysis sued the entire feature set for each emotion detection system (i.e., API). The results demonstrate that by including additional features, all APIs yielded significantly greater predictive accuracy. For example, the best performing model (Microsoft Cognitive Services) achieved a test accuracy of 82.89% and a test Kappa of 76.81% (mean CV accuracy of 82.01% and mean CV Kappa of 75.77% across all emotion detection APIs).

Incorporating additional features increased the predictive power also of negative discrete emotions. Yet, we still observed systematically greater variation in accuracy measures for negative compared to positive discrete emotions ($\sigma_{\text{happy open}} = 3.84$, $\sigma_{\text{neutral}} = 9.96$, $\sigma_{\text{fear}} = 21.54$, $\sigma_{\text{anger}} = 15.5$).

6. Accuracy Improvement 1: Model Averaging Across APIs & Feature Combination

How can we further improve the predictive accuracy of a model? One option is to increase the feature space by combining the entire set of features across all APIs. The second option is to use more flexible estimation techniques. Specifically, recent advances in deep learning might further improve predictive performance by increasing the number of layers and / or boosting a model by aiming at predicting misclassified cases. With respect to the first option of combining feature sets, our results highlight that the obtained test accuracy improved further with a mean CV accuracy of 85.73% and mean CV Kappa of 81.02%.

7. Accuracy Improvement 2: Flexible Estimation Techniques

Next, we aimed at further improving prediction accuracy by using more flexible estimation techniques. Parameters were selected through grid search in the parameter space. Lowest performance was achieved by Support Vector Machines with radial basis kernel resulting in a test accuracy of 84.41% and a test Kappa of 79.33% (mean CV accuracy of 83.47% and mean CV Kappa of 78.05%). Highest performance was achieved by using Random Forests and LogitBoost achieving 92.02% test accuracy, 89.25% test kappa (mean CV accuracy of 90.75% and mean CV kappa of 87.62%) and 93.60% accuracy, 91.34% kappa (mean CV accuracy of 90.88% and mean CV kappa of 87.62%), respectively.

This modeling strategy led to a significant increase in predictive power also for negative discrete emotions. The predictive power difference between positive, neutral and negative classes was reduced ($\sigma_{\text{happy open}} = 2.71$, $\sigma_{\text{neutral}} = 1.36$, $\sigma_{\text{fear}} = 5.32$, $\sigma_{\text{anger}} = 4.61$).

8. General Discussion

The current work makes three novel contributions. First, we demonstrate that current emotion detection systems yield substantive variation in predictive accuracy across discrete emotions. Second, we demonstrate two modeling strategies to increase predictive accuracy both overall and specifically for target emotions. The first strategy is to expand the feature set by combining features across emotion detection systems. The second strategy is to use more flexible modeling techniques aiming at minimizing prediction error for misclassified classes and by using ensemble methods. Specifically, we show that ensembles of decision trees outperform a variety of other machine learning models (in terms of cross-validated accuracy on the test set).

To the best of our knowledge, this is the first systematic study demonstrating the striking variability in predictive accuracy of automated emotion detection systems across discrete emotions and we provide two easy to implement modeling strategies to counter these inaccuracies.

9. References

Bashyal, Shishir and Ganesh K Venayagamoorthy (2008), “Recognition of facial expressions using Gabor wavelets and learning vector quantization,” *Engineering Applications of Artificial Intelligence*, 21 (7), 1056–64.

Cottrell, Garrison W and Janet Metcalfe (1991), “EMPATH: Face, emotion, and gender recognition using holons,” in *Advances in neural information processing systems*, 564–71.

Kanade, Takeo (1973), “Picture processing by computer complex and recognition of human faces,” *Ph. D. Thesis, Kyoto University*.

Levi, Gil and Tal Hassner (2015), “Emotion recognition in the wild via convolutional neural networks and mapped binary patterns,” in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 503–10.

Littlewort, G C, Marian Stewart Bartlett, Ian R Fasel, Joel Chenu, Takayuki Kanda, Hiroshi Ishiguro, and Javier R Movellan (2004), “Towards social robots: Automatic evaluation of human-robot interaction by facial expression classification,” in *Advances in neural information processing systems*, 1563–70.

Liu, Ping, Shizhong Han, Zibo Meng, and Yan Tong (2014), “Facial expression recognition via a boosted deep belief network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1805–12.

Ma, Debbie S, Joshua Correll, and Bernd Wittenbrink (2015), “The Chicago face database: A free stimulus set of faces and norming data,” *Behavior research methods*, 47 (4), 1122–35.

Padgett, Curtis and Garrison W Cottrell (1997), “Representing face images for emotion classification,” in *Advances in neural information processing systems*, 894–900.

Rani, Pramila, Changchun Liu, Nilanjan Sarkar, and Eric Vanman (2006), “An empirical study of machine learning techniques for affect recognition in human--robot interaction,” *Pattern Analysis and Applications*, 9 (1), 58–69.

Yu, Zhiding and Cha Zhang (2015), “Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 435–42.