Racial Influence on Automated Perceptions of Emotions

Lauren Rhue
rhuela@wfu.edu

Introduction

The practical applications of artificial intelligence are expanding into various elements of society, leading to a growing interest in the potential biases of such algorithms. Facial analysis, one application of artificial intelligence, is increasingly used in real-word situations like human resources and threat prediction. In human resources, some organizations use hiring video platforms screen their candidates. These candidates answer predefined questions in a recorded video, and organizations can use facial recognition to analyze the potential applicant faces (Zetlin, 2018). The analysis influence the hiring process, such as whether the manager ever sees the video.

In threat prediction, companies are developing facial recognition software to scan the faces in crowds and assess if any of the individuals pose a public safety threat. One company, WeSee, claims to assess the person's "mental state" using cues that are imperceptible to the human eye (Thomas, 2018). This application specifically mentions emotions such as doubt and anger as emotions that indicate threats.

In both these situations, facial recognition can pose notable consequences on individuals. If an AI system mistakenly views a candidate as angry, then the person may never receive a call-back interview or find a position in their field. If an AI system identifies an individual as a threat, then that person could be detained, followed, placed on a no-fly list, or some other significant consequence. A false arrest may haunt a person for years and reduce their employability. Given the potentially life-altering consequences of facial recognition AI, the research community should consider the potential bias in such systems.

This study provides evidence that facial recognition software interprets emotions differently based on the person's race. Using a publically available data set of professional basketball players' pictures, I compare the emotional analysis from two different facial recognition services, Face++ and Microsoft AI. Both services interpret black players as having more negative emotions than white players; however, they present bias in two different ways. Face++ consistently interprets black players as angrier than white players, even controlling for their degree of smiling. Microsoft registers contempt instead of anger, and it interprets black players as more contemptuous when their facial expression is ambiguous. As the players' smile widens, the disparity disappears.

Background

Research has extensively shown the presence of racial discrimination in technical systems. In particular, outcomes differ based on whether the person has lighter or darker skin. An eBay auction with a dark-skinned hand holding the product receives lower-priced bids than another auction with a white hand holding the same product (Ayres et al., 2015). On Kickstarter, contributors discount the value of products from black entrepreneurs (Younkin and Kuppuswamy, 2018) and black entrepreneurs experience as 50% lower success rate than other racial groups (Rhue and Clark,

2018). Parker and Meija (2018) also find the presence of increased driver cancelations for darker skin riders. Numerous studies have confirmed the presence of racial discrimination in technical systems.

The research literature increasingly acknowledges that algorithms are not neutral (O'Neil, 2016). The use of big data can thus lead to artificial intelligence to produce a disparate impact for minorities, women, and/or other traditionally disadvantaged group (Barocas and Selbst, 2014). For example, Boston's Street Bump mobile application automatically detected potholes and alerted the city; however, Street Bump detected more potholes in affluent neighborhoods because lower-income constituents were less likely to own smart phones and download Street Bump (Crawford, 2013). Seemingly innocuous algorithms, such as scheduling algorithms to respond to customer demand, can exacerbate existing power differences and harm workers (Barocas and Levy, 2016). Buolamwini and Gebru (2018) found that facial recognition software predicts gender significantly worse for darker-skin faces than lighter-skin faces.

With the increased evidence of bias in artificial intelligence, this study expects to find that facial recognition software interprets facial expressions differently by race and assigns more negative emotion to black faces.

Empirical Analysis

This study uses a publicly available data set of all NBA players from the 2016-17 season, more than 400 faces, from Basketball Reference[1]. NBA players are relatively homogenous in their age, gender, and physicality. Furthermore, these pictures are taken in a professional context with relatively standard poses. All players face the camera head-on, a preferable position for accurate facial analysis.

I score the pictures using two different facial recognition software options that analyze the facial expression for emotions: Face++[2] and Microsoft AI[3]. Other popular facial recognition services such as IBM Watson did not offer an emotional analysis.

Face++ analyzes faces for emotion, smile, gender, race, and quality. For emotion, each face is assigned a value on each of the seven emotions of Anger, Disgust, Fear, Happy, Neutral, Sadness, and Surprise. The sum of the seven emotional scores equals 100. Face++ identifies the degree to which a face is smiling, and this attribute is distinct from the assessment of happiness. Face++ also includes a continuous measure of facial quality to indicate the degree of noise in the data.

Microsoft analyzes faces on emotion, smiling, and noise level. Each face is scored on eight emotional dimensions: Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. For each face, all the emotions sum to 1. In addition to emotions, Microsoft rates the smile between 0-1, and the degree of the smile is the same score for the face's Happiness. Microsoft also includes a categorical measure of noise: Noise (Low), Noise (Medium), and Noise (High).

[1] https://www.basketball-reference.com/
[2] https://www.faceplusplus.com/
[3] https://www.microsoft.com/en-us/ai

Summary Statistics

In the initial pass, there are notable differences between the average emotional analyses for players by race.

INSERT TABLE 1

Microsoft's AI only registers five players as having anger and interprets the anger of black and white players similarly. However, Microsoft includes an additional emotional category of contempt. On this emotional dimension, black players are viewed as more than 3x as contemptuous as white players. Face++ interprets black players has more than 2x as anger as white players. Disgust is insignificant between the two populations, and fear is more than 3x higher for black players. Microsoft AI views black and white players as equally happy, but Face++ interprets black players as 20% less happy. Microsoft and Face++ systems viewed both populations as similarly neutral.

For some emotions, black and white players are indistinguishable; however, the AIs appear more prone to assign negative emotions to black players. To examine this conjecture, I compare the "positive" emotion of happiness and the "negative" emotions of anger and contempt.

INSERT FIGURE 1

Figure 1a shows that both Microsoft and Face++ assign black players a lower level of happiness than white players. Figure 1b shows that Face++ rates black players as significantly higher in angry. Microsoft does not register much anger for either players but scores black players as more contemptuous.

Alone, this observation does not indicate that the AIs are systematically bias in their interpretation. Black players could indeed have consistently angrier facial expressions than white players, so AIs could be accurately assessing their emotions. To account for this possibility, I control for the degree to which the players are smiling. If the AI recognizes a similar degree of smiles from the players yet interprets the emotions differently, there is evidence that AIs assign more negative emotions to black players than white players.

Matched Sample

To examine whether AIs assign negative emotions to black players, I create a quasi-experimental situation by matching black and white players based on the degree of their smile and other characteristics. If black players are scored with more negative emotions than white players, even controlling for the degree of smiling, then there is evidence that AIs assign more negative emotion to black players.

The following example demonstrates the principle of matched samples. I choose two players, Darren Collison and Gordon Hayward. As shown in Figure 2, both men are smiling somewhat although Collison is smiling with his mouth open and Hayward is smiling with his mouth closed.

INSERT FIGURE 2

According to Face++, Darren Collison and Gordon Hayward have similar smile scores of 48.7 and 48.1 respectively. Translated into emotion, Face++ rates Hayward's expression as 59.7% happy

and 0.13% angry yet rates Collison's expression as 39.2% happy and 27% angry. Collison is rated as more than 180x angrier than Hayward despite his smile!

Microsoft's AI is more accurate, rating both men as primarily happy, but a gap remains between the two players. Collison is scored as 5 percentage less happy than Hayward (98% and 93% respectively) so Collison is perceived as less happy than Hayward. Despite a wide smile, Collison even has a negligible amount of contempt associated with his facial expression (0.1%) whereas Hayward has none (0%).

This example demonstrates the principle behind matched sample techniques. To create the matched sample, I employ Coarsened Exact Matching (CEM). I stratify the smile variable into 10 bins and match black and white players with the same degree of smile. Unmatched players are removed from the analysis.

Estimation Results

Face++

The model specification controls for the available elements that could influence the emotional analysis, such as the noise and age. For Face++, the negative emotion of interest is *Anger*. There are two different models of Anger. First, *Anger* is modeled as a continuous score, indicating the degree to which a black player is perceived as angrier. Second, *AngerIndicator* is a binary, indicating whether a black player is more likely to be perceived as having more than residual anger.

$$Anger = \beta_1 Black + \beta_2 Smile + \beta_3 Noise + \beta_4 Covariates + \epsilon \quad (1)$$

$$AngerIndicator = \beta_1 Black + \beta_2 Smile + \beta_3 Noise + \beta_4 Covariates + \epsilon \quad (2)$$

In equation (1), the dependent variable is *Anger*, the Face++ anger score; in equation (2), it is *AngerIndicator*, a binary indicator for whether Anger is positive. For equations (1) and (2), the independent variables are the same. *Black* is a binary indicator for whether the player is black. *Smile* is the degree of the smile associated with the player's face. *Noise* is the continuous score of the face quality according to Face++, and covariates are the *Age* and player's home *Country*.

INSERT TABLE 2

Table 2 shows the results of this analysis. The first three columns show the OLS estimation results for equation (1). The coefficient estimate on *Black* is 3.58, suggesting that, all else being equal, black players are more than twice as likely to be interpreted as angry. As expected, *Smile* is significant and negative, showing that each unit increase in the smile degree is associated with a 0.08 unit decrease in the associated anger. *Face Quality* and *Age* are not significantly associated with predicted anger.

The last three columns show the logistic regression estimates with the dependent variable Anger > 1. The coefficient estimate of *AngerIndicator* is approximately 1.1, providing additional evidence that black players are more likely to receive higher *Anger* scores. Smile is negative, showing that each unit increase in smile decreases the log odds of being perceived as angry.

The Face++ results are fairly clear and consistent. Black players score as incrementally angrier for any degree of smiling, and black players are more likely to be perceived as angry.

Microsoft

Next, this study uses a similar model to analyze Microsoft's AI. For Microsoft, the negative emotion of interest is *Contempt*.

$$Contempt = \beta_1 Black + \beta_2 Smile + \beta_3 Noise + \beta_4 Covariates + \epsilon \quad (3)$$

$$ContemptIndicator = \beta_1 Black + \beta_2 Smile + \beta_3 Noise + \beta_4 Covariates + \epsilon \quad (4)$$

In equation (3), the dependent variable is *Contempt*, the Microsoft contempt score; in equation (2), it is *ContemptIndicator*, a binary indicator for whether *Contempt* is positive. For equations (3) and (4), the independent variables are the same as above. *Black* is a binary indicator for whether the player's race is black. *Smile* is the degree of the smile associated with the player's face. *Noise* is the categorization of the noise according to Microsoft (Low, Medium, and High), and covariates are the *Age* and player's home *Country*.

INSERT TABLE 3

Table 3 shows the coefficient estimates. The first three columns show the OLS estimates for equation (3), and there are no significant difference between the perceived contempt for black and white players. Surprisingly, even the coefficient estimate for Smile is not significant in the matched sample. This model does not appear to fit the data. The last three columns show the logistic regression estimates for equation (4). The coefficient estimate for *Black* is significant and positive, suggesting that the AI is more likely to score black players as contemptuous. The coefficient estimate for *Smile* is significant and negative so wider smiles are associated with lower log odds of a positive score for contempt.

The Microsoft results are not as clear or as consistent as the Face++ results. Although black players are more likely to be perceived as contemptuous, the relationship is not linear. To further interpret the results, I plot the negative emotion (Anger or Contempt) as a function of the Smile for Face++ and Microsoft. Then I compare the fitted models for black and white players. These visualizations suggest two different mechanisms for AI bias.

Mechanisms

The plots suggest two distinct mechanisms that would both lead to an increase in average bias. First, AIs could display a consistent bias in their results and assign black faces a higher negative emotion. Second, AIs could interpret ambiguous facial expressions more negatively for black faces. Essentially, AIs would not give black faces "the benefit of the doubt" in uncertain situations. There is evidence that these mechanisms are behind the results.

Consistent Bias

As shown in Figure 3, the fitted model for black players is consistently higher than the model for white players until *Smile* surpasses 90.

INSERT FIGURE 3

Thus, Face++ assigns higher scores for black faces at every level of Smile. This is a consistent interpretation of black players as angrier than white players.

Interpretation of Ambiguity
Microsoft assigns more negative emotions in ambiguous situations but the difference between black and white players disappears as the players' smile scores approach unity.

INSERT FIGURE 4

As shown in Figure 4, the fitted model for black players is consistently higher than the model for white players when Smile is greater than 0 and less than 0.75. For the extremes, the Microsoft AI interprets the emotions of black and white players similarly and the disparity disappears.

Implications

AIs display racial disparities in their emotional scores and are more likely to assign negative emotion to black men's faces. Face++ interprets black players as angrier for every level of *Smile*. Microsoft only interprets black players as more contemptuous for ambiguous and/or non-smiling pictures. The analysis controls for facial quality, so this finding is not a result of the pictures themselves.

This paper has some limitations. First, this paper finds the presence of racial disparities in the emotional scores, but are AIs better at accurately deciding emotion than people? Perhaps the AI determines emotion more accurately than people do. Second, this analysis focuses only on men so it does not cover the gender or racial differences in emotional scores. Third, the analysis focuses perception of anger and contempt because the potential for negative consequences, but there are other emotions.

These results add to the growing literature on fairness in AI and have implications for the artificial intelligence community, individuals, and organizations using AI. First, AI developers must continue to refine their models and analyze them for disparate impact (Barocas and Levy, 2016). Facial recognition programs are not uniform, so the bias may be introduced through different mechanisms such as a consistent bias or a bias in ambiguous facial expressions.

Second, professionals of color should exaggerate their facial expressions –smile more—to reduce the potential negative interpretations. Is this additional burden fair? No. However, this recommendation to smile more aligns with Grandey et al. (2018) who observe that black service providers need to amplify positive emotions in order to receive parity in their evaluations. Of course, this findings places additional emotional burden on black professionals.

Third, organizations who use AI to screen candidates should monitor and review the results before overly relying on the system. If professionals of color are systematically viewed as having more negative emotions, then they could be eliminated from the interview pool prematurely. AIs could lead to disproportionate impact for candidates of color, which violates Equal Opportunity regulations, hurts diversity / inclusion effects, and eliminates good candidates.

Overall, this research builds upon the emerging literature on racial differences in artificial intelligence and provides evidence that the person's race influences how facial recognition software interprets facial expressions.

References

Ayres, I., Banaji, M., & Jolls, C. (2015). Race effects on eBay. *The RAND Journal of Economics*, *46*(4), 891-917.

Barocas, S. and Levy, K. (2016). What Customer Data Collection Could Mean for Workers. *Harvard Business Review.* Available at: *https://hbr.org/2016/08/the-unintended-consequence-of-customer-data-collection*

Barocas, S., and Selbst, A.D. (2014). Big Data's Disparate Impact. *California Law Review* 104 (3), 671-732

Buolamwini, Joy, and Timnit Gebru. (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on Fairness, Accountability and Transparency. 2018.

Crawford, K. (2013). The Hidden Biases in Big Data. *Harvard Business Review*, available at: https://hbr.org/2013/04/the-hidden-biases-in-big-data.

Edelman, B., Luca, M., and Svirsky, D. (2017). "Racial discrimination in the sharing economy: Evidence from a field experiment." *American Economic Journal: Applied Economics* 9(2) (2017): 1-22.

Grandey, A., Houston III, L., and Avery, D. (2018). Fake It to Make It? Emotional Labor Reduces the Racial Disparity in Service Performance Judgments. *Journal of Management.* Vol. XX No. X, Month XXXX 1–30.

Mejia, J., & Parker, C. (2018). When Transparency Fails: Bias and Financial Incentives in Ridesharing Platforms. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3209274.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, New York, NY, USA.

Thomas, D. (2018). The cameras that know if you're happy - or a threat. BBC News, July 17, 2018. Available at: https://www.bbc.com/news/business-44799239.

Rhue, L. and Clark, J. (2018). "The Consequences of Authenticity: Quantifying Racial Signals and their Effects on Crowdfunding Success." Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2837042

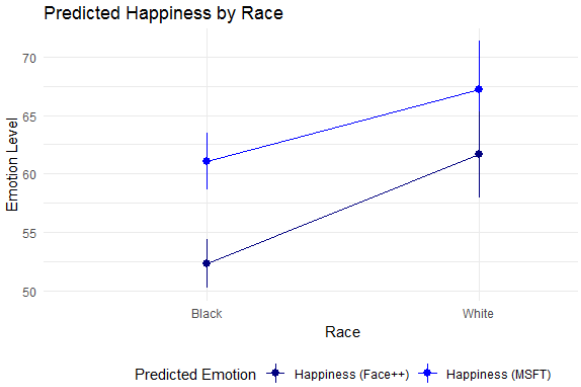
Younkin, P., and Kuppuswamy, V. (2017). "The Colorblind Crowd: Founder Race and Performance in Crowdfunding," Management Science, *forthcoming*.

Younkin, P., & Kuppuswamy, V. (2018). Discounted: The effect of founder race on the price of new products. *Journal of Business Venturing, forthcoming*.

Zetlin, M. (2018). AI Is Now Analyzing Candidates' Facial Expressions During Video Job Interviews. Inc.com. https://www.inc.com/minda-zetlin/ai-is-now-analyzing-candidates-facial-expressions-during-video-job-interviews.html. Retrieved Nov 5, 2018.

Figures

Figure 1. Initial Emotional Comparison



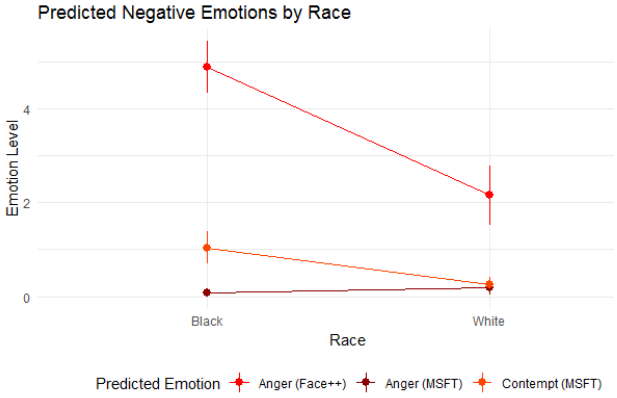


Figure 2. Example Pictures

Figure 2. Darren Collison (L), Gordon Hayward (R) (Source Basketball Reference)

Figure 3. Relationship between Anger and Smiling

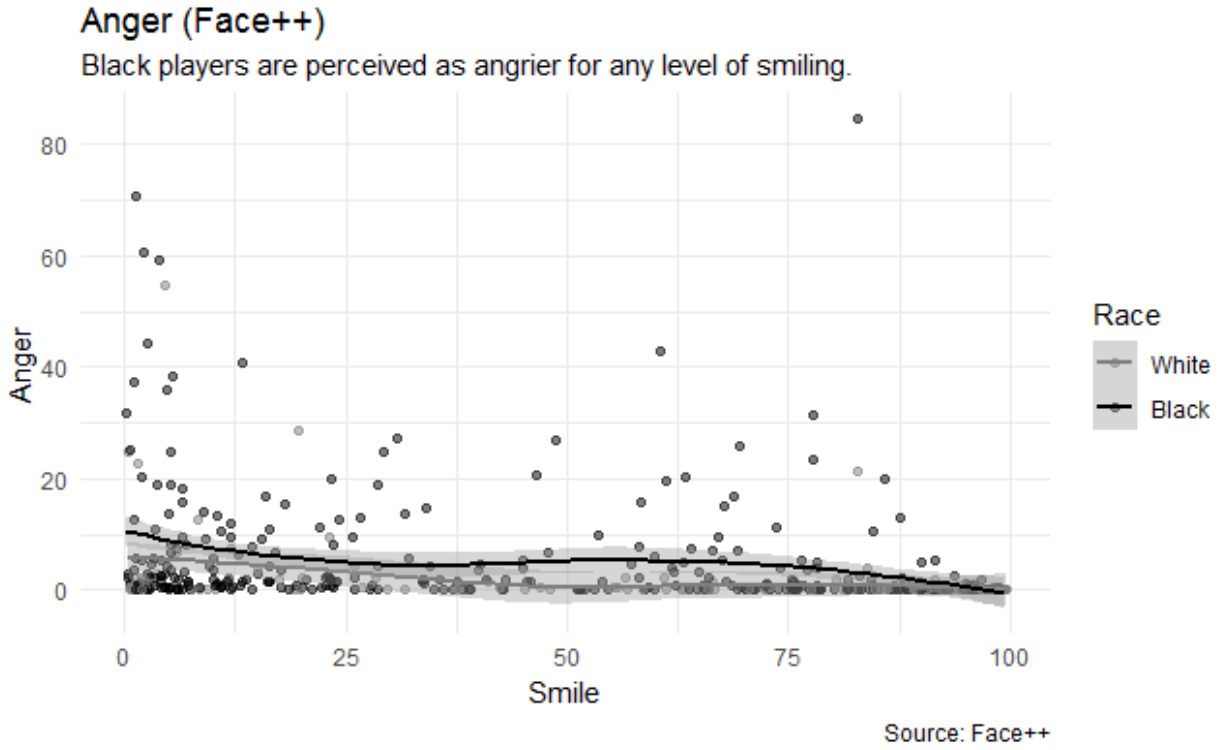
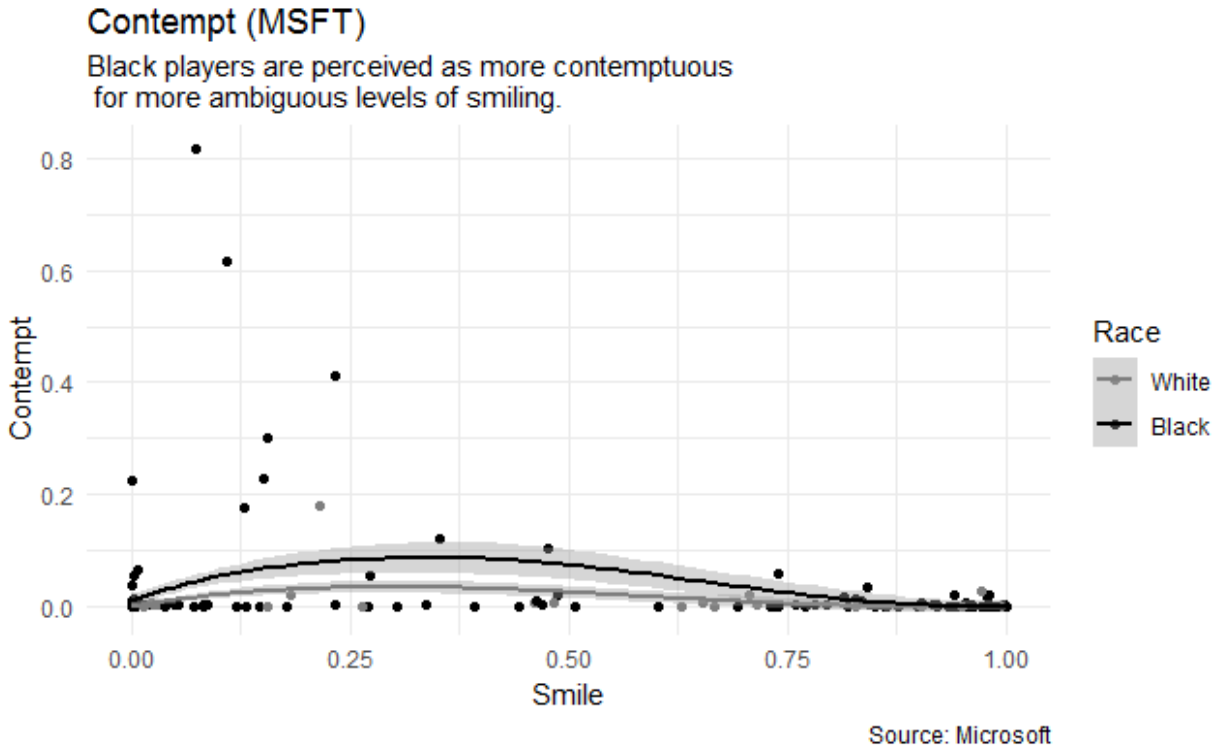


Figure 4. Relationship between Contempt and Smiling (Microsoft)



Tables

Table 1. Average Emotional Analysis by Race

	Microsoft				Face++			
	Black		White		Black		White	
Emotion	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Anger	0.001	0.000	0.002	0.002	4.880	0.546	2.149	0.628
Contempt	0.010	0.003	0.003	0.002	N/A		N/A	
Disgust	0.000	0.000	0.000	0.000	10.967	0.985	10.299	1.817
Fear	0.000	0.000	0.000	0.000	4.839	0.609	1.156	0.239
Happy	0.611	0.024	0.672	0.041	52.301	2.095	61.654	3.671
Neutral	0.373	0.024	0.322	0.041	17.749	1.376	19.198	2.936
Surprise	0.001	0.000	0.000	0.000	7.401	0.633	3.770	0.990
Sadness	0.004	0.002	0.001	0.000	1.863	0.321	1.772	0.351
Smile	0.611	0.024	0.672	0.041	50.189	1.884	56.674	3.361

Table 2. Face++ Emotional Analysis

	Dependent Variable: Anger			*Dependent Variable: AngerIndicator*		
	1	2	3	4	5	6
Variable	Estimate (Std. Error)	Estimate (Std. Error)	Estimate (Std. Error)	Estimate (Std. Error)	Estimate (Std. Error)	Estimate (Std. Error)
Intercept	4.483	-3.202	-4.955	0.998	-1.023	0.104
	(18.127)	(18.935)	(8.272)	(4.630)	(4.919)	(2.136)
Black	3.529	3.576	2.492	1.122	1.146	1.080
	(1.301)**	(1.302)**	(1.172)*	(0.360)**	(0.363)**	(0.334)**
Smile	-0.082	-0.080	-0.075	-0.025	-0.025	-0.029
	(0.015)***	(0.015)***	(0.013)***	(0.004)***	(0.004)***	(0.003)***
Face	0.016	0.024	0.044	-0.015	-0.009	0.005
Quality	(0.197)	(0.198)	(0.051)	(0.050)	(0.051)	(0.014)
Current		0.228	0.126		0.037	0.032
Age		(0.168)	(0.104)		(0.044)	(0.027)
Origin Country			Yes			Yes
US Indicator		Yes			Yes	
Sample	Matched Sample	Matched Sample	All		Matched Sample	Matched Sample	All
R^2	0.132	0.14	0.086	AIC	274.34	276	542.37
N	258	258	471	N	258	258	471

Table 3. Microsoft Emotional Analysis

	Contempt			*Contempt Indicator*		
	1	2	3	4	5	6
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Variable	(Std. Error)	(Std. Error)	(Std. Error)	(Std. Error)	(Std. Error)	(Std. Error)
Intercept	0.314	0.221	0.010	-0.714	-0.693	-0.586
	(0.777)	(1.798)	(0.020)	(0.444)	(1.115)	(0.784)
Black	0.419	0.418	0.007	0.726	0.727	0.364
	(0.498)	(0.499)	(0.007)	(0.345)*	(0.346)*	(0.279)
Smile	-0.008	-0.008	-0.017	-0.013	-0.013	-1.386
	(0.005)	(0.005)	(0.006)**	(0.003)***	(0.003)***	(0.227)***
Noise –	0.038	0.057	-0.002	-0.733	-0.740	-0.282
Low	(0.717)	(0.720)	(0.008)	(0.399)	(0.401)	(0.315)
Noise –	0.683	0.710	0.002	-0.470	-0.479	-0.383
Medium	(0.739)	(0.745)	(0.008)	(0.405)	(0.409)	(0.327)
Current Age		-0.008	0.000		0.003	0.024
		(0.056)	(0.001)		(0.036)	(0.024)
US Country		0.323	0.000		-0.101	-0.182
		(0.707)	(0.007)		(0.434)	(0.284)
Sample	Matched	Matched	All	Matched	Matched	All
R^2	0.015	0.016	0.01	334.1997	338.4374	544.61