TABLE II. AUTHOR RACE/ETHNICITY CLASSIFICATION PERFORMANCE MEASURES

Race/Ethnicity Class	Ad n	Algorithm	Recall μ (SD)	Precision μ (SD)	F_1 score μ (SD)	vs. Baseline F_1 score
Biracial	14	*First-Mention*	0.143 (0.00)	0.933 (0.13)	0.247 (0.01)	-
		CRF-First	0.093 (0.03) *	**1.000 (0.00) ***	0.168 (0.05) *	-31.9%
		CRF-All	**0.600 (0.03) ***	0.672 (0.05)*	**0.632 (0.02) ***	155.8%
Hispanic/Latino	34	*First-Mention*	**0.853 (0.00)**	0.706 (0.01)	0.772 (0.00)	-
		CRF-First	0.829 (0.01) *	0.788 (0.01) *	**0.808 (0.01) ***	4.6%
		CRF-All	0.724 (0.01) *	**0.886 (0.03) ***	0.796 (0.01) *	3.1%
Black	27	*First-Mention*	**1.000 (0.00)**	0.565 (0.01)	0.722 (0.01)	-
		CRF-First	0.970 (0.02) *	0.688 (0.02) *	0.805 (0.01) *	11.5%
		CRF-All	0.941 (0.03) *	**0.782 (0.02) ***	**0.854 (0.02) ***	18.3%
Asian	19	*First-Mention*	0.900 (0.02)	0.740 (0.00)	0.812 (0.01)	-
		CRF-First	**0.947 (0.00)***	0.896 (0.01) *	**0.921 (0.01) ***	13.4%
		CRF-All	0.879 (0.02) *	**0.943 (0.00) ***	0.910 (0.01) *	12.0%
Caucasian	136	*First-Mention*	0.868 (0.00)	0.851 (0.01)	0.859 (0.00)	-
		CRF-First	**0.909 (0.01) ***	0.920 (0.01) *	0.914 (0.01) *	6.4%
		CRF-All	0.900 (0.01) *	**0.935 (0.01) ***	**0.917 (0.01) ***	6.8%
Undisclosed	299	*First-Mention*	0.874 (0.00)	**0.952 (0.00)**	0.912 (0.00)	-
		CRF-First	**0.948 (0.01) ***	0.941 (0.00) *	**0.944 (0.00) ***	3.6%
		CRF-All	**0.948 (0.01) ***	0.916 (0.00) *	0.932 (0.00) *	2.2%
Hawaiian/Pacific-Islander	3	*First-Mention*	**0.333 (0.00)**	0.228 (0.04)	0.269 (0.03)	-
		CRF-First	**0.333 (0.00) ***	**1.000 (0.00) ***	**0.500 (0.00) ***	85.8%
		CRF-All	**0.333 (0.00) ***	**1.000 (0.00) ***	**0.500 (0.00) ***	85.8%

GOLD race/ethnicity classification performance measures. *First-Mention* is a simple rule-based heuristic which assigns author race using the first race/ethnicity term identified in ad text. *CRF-First* and *CRF-All* are hybrid machine learning/rule-based approaches that generate labels to identify author race. Both *CRF-First* and *CRF-All* use *First-Mention* as a baseline for percentage improvement in F-score, with * indicating a statistically significant difference ($p < 0.01$) using a paired t-test. Overall, *CRF-All* provides significant improvements over the baseline, especially in precision, This method performs poorly in classifying Hawaiian/Pacific-Islanders due to low sample size and in identifying Biracial ads, largely because of terminology overlap with other classes (e.g., "I am a hispanic/white male.")

TABLE III. CRAIGSLIST RACE/ETHNICITY VS. 2010 CENSUS

Race/Ethnicity Class	$\lceil$Ad $m\rceil$	Geounit Type	Geounit n	CRAIGSLIST Corpus			PHONE Corpus		
				R^2	Coef.	[95% CI]	R^2	Coef.	[95% CI]
Biracial	365,811	Metropolitan	365	0.43	0.82	[0.72, 0.92]	0.26	0.73	[0.60, 0.86]
	13,868	Micropolitan	523	0.53	0.60	[0.55, 0.64]	0.32	0.70	[0.62, 0.79]
	30,576	County	587	0.76	0.76	[0.73, 0.79]	0.47	0.72	[0.67, 0.77)
Hispanic/Latino	963,810	Metropolitan	365	0.91	0.81	[0.78, 0.84]	0.81	0.81	[0.77, 0.85]
	19,833	Micropolitan	523	0.56	0.54	[0.50, 0.58]	0.25	0.58	[0.50, 0.66]
	36,853	County	587	0.66	0.61	[0.58, 0.64]	0.54	0.71	[0.66, 0.75]
Black	992,014	Metropolitan	365	0.71	0.36	[0.34, 0.39]	0.54	0.46	[0.41, 0.5]
	38,985	Micropolitan	523	0.44	0.33	[0.30, 0.37]	0.24	0.51	[0.44, 0.59]
	41,675	County	587	0.45	0.24	[0.22, 0.26]	0.19	0.30	[0.26, 0.34]
Asian	359,114	Metropolitan	365	0.86	0.65	[0.63, 0.68]	0.63	0.57	[0.52, 0.61]
	11,472	Micropolitan	523	0.54	0.55	[0.51, 0.60]	0.36	0.58	[0.51, 0.64]
	38,535	County	587	0.66	0.55	[0.53, 0.58]	0.44	0.44	[0.41, 0.47]
Caucasian	4,869,614	Metropolitan	365	0.04	-0.24	[-0.35, -0.13]	0.04	-0.24	[-0.36, -0.11]
	215,777	Micropolitan	523	0.05	-0.38	[-0.52, -0.24]	0.01	-0.27	[-0.52, -0.02]
	218,246	County	587	0.04	-0.18	[-0.23, -0.12]	0.10	-0.46	[-0.55, -0.37]
Undisclosed (vs. Caucasian Census)	18,349,737	Metropolitan	365	0.34	0.23	[0.20, 0.27]	0.39	0.31	[0.27, 0.35]
	1,006,751	Micropolitan	523	0.19	0.21	[0.17, 0.24]	0.08	0.36	[0.26, 0.45]
	962,799	County	587	0.23	0.13	[0.11, 0.14]	0.07	0.19	[0.15, 0.24]
Native Hawaiian/Pacific Islander	20,289	Metropolitan	365	0.42	0.40	[0.35, 0.45]	0.21	0.40	[0.32, 0.47]
	1,646	Micropolitan	523	0.94	0.37	[0.37, 0.38]	0.90	0.47	[0.46, 0.48]
	4,237	County	587	0.96	0.45	[0.44, 0.45]	0.88	0.55	[0.54, 0.56]

CRAIGSLIST OLS log-log regression results for census race/ethnicity percentages (independent variable) vs. Craigslist race/ethnicity disclosures per 100 MSM ads (dependent variable). Data points consist of the percentage population makeup, as determined by 2010 census data and measured Craigslist disclosures. Ad m (which can take fractional weights when crossing multiple geographic boundaries) is rounded up. Coefficients are in log units meaning, for example, a 1% increase in census population makeup for the Hispanic/Latino group results in a 0.81% C.I [0.78, 0.84] increase in Craigslist ads disclosing Hispanic/Latino origin.

B. Author Age

In the GOLD corpus, 90% (627/700) of ads contained author age information. Our regular expression correctly matched 95% (594/627) of all age tags with precision 0.99, recall 0.95 and F_1 score 0.97. Age correlations were strongest in metropolitan geographic areas, with the highest correlations