attempt at disambiguation is attempted in this approach, which provides our baseline performance measure.

1	5 ' 9 ' 140 **blk** brn 30w 7c . . mix **blk** / **mexican**	
	NNNN N **N** N N N NN N **A** N **A**	
2	all race welcom . . . **latino** is a plus	
	N N N NNN **P** N N N	

Fig. 2. Excerpt of labeled output. Each term in a sequence is assigned a `label` ∈ { `AUTHOR (A), PARTNER (P), NONE (N)`} predicted based on 13 features, using conditional random fields. Given this labeling, both *First Mention* and *CRF-First* would incorrectly classify this ad's author as Black. *First Mention* fails to disambiguate the first usage (as hair color) of "blk" while *CRF-First* only considers the first race term labeled as `AUTHOR` . *CRF-All*, which considers all `AUTHOR` labels, would correctly predict Biracial.

2) Hybrid Method: This approach uses 11 Boolean and 2 nominal features (see below) to assign each term a `label` ∈ {`AUTHOR, PARTNER, NONE`} using linear-chain conditional random fields (CRFs). CRFs are undirected graphical models that, in the special case of a linear chain graph structure, can be used to efficiently label sequence data [32]. This machine generated label set is used by a rule-based classifier to then assign an ad to one of the 6 possible race/ethnicity categories (or Undisclosed if no `AUTHOR` tags are found or the labeled term isn't in our thesaurus). By constraining the machine learning step to detect all race/ethnicity mentions, independent of the class that mention belongs to, we help prevent overfitting in the less frequent categories in our training corpus.

For the rule-based classification, we consider two variations of the *First Mention* rule discussed above; (1) *CRF-First*; and (2) *CRF-All*. *CRF-First* uses the first `AUTHOR` labeled term in text to predict a race/ethnicity category, not just the first observed race/ethnicity vocabulary term. Second, *CRF-All* consider the set of all `AUTHOR` labels when assigning a category. If an ad contains `AUTHOR` labeled terms from more than one race/ethnicity terminology cluster, it is assigned to the Biracial class if those terms are separated by a slash (e.g., "white / black") and are not contained within a list. See Figure 2 for an example labeling and its resulting classification.

The performance of these methods is evaluated on the annotation corpus, averaged over 10 trials, with each trial using stratified, 10-fold cross validation. To build our race/ethnicity terminology lexicons, the n-1 folds of annotated training data are used to create the thesaurus used in labeling the documents of the nth fold. This cross-validation approach ensures that we capture the effects of lexical acquisition in our classifier. For *CRF-First* and *CRF-All*, stemmed token word windows of size 3-9 were tested with all features, with 6 performing best overall. Other features such as part-of-speech tags were tested, but did not result in a statistically significant improvement in performance and were not included in the final feature set below:

- *Stemmed Term (Nominal)*: The current term and a 6-term window of all surrounding words.
- *Race/Ethnicity Category (Nominal)*: Name of this term's parent race/ethnicity terminology cluster or None otherwise.
- *Digit + Unit of Measurement*: Term is a unit of measurement, e.g., "5'8'" "130lbs."
- *Metadata*: Term is part of age, location, or encounter tag metadata.
- *First Mention*: Term is the first labeled race/ethnicity term in ad text.
- *List*: Term co-occurs within a 5-term window of other race/ethnicity mentions, e.g., "totally into black, asian, latin or ethnic guys."
- *Pluralization*: Term belongs to a race/ethnicity category and ends in "s".
- *Partner Preference*: Qualifying terms within a 5-term window that express negation or partner preference, i.e., `term` ∈ {`no, not, into, none, only`}.
- *Punctuation*: Term is a punctuation mark.
- *Slash*: Term is in a race/ethnicity category and is separated from another race/ethnicity term by a slash character.
- *Left Verb Argument*: Term is within a 5-term left-window of a set of left/right-associative verbs: `v` ∈ {`looking, seeking, wanted`}. The verb "looking" is ignored if it occurs in the bigrams "good looking" or "nice looking."
- *Right Verb Argument*: Same as above but for the right-term window.
- *PRP + Being Verb*: Term is preceded by a syntactic pattern of the form `PersonalPronoun + BeingVerb` (e.g., "I am") where `BeingVerb` ∈ {`am,'m,:`}.

D. Calculating Demographic Rates

Using the methods described thus far, we extract race/ethnicity and age from all MSM ads. Demographic prevalence rates are calculated by collapsing all ads associated with a geographic region into a single bin L and checking for ads containing search terms associated with race/ethnicity and age attributes. Prevalence rate is then simply the percentage of MSM ads containing the search terms in question for a given location and time interval.

For our demographic analysis, L is defined as the set of all location tag toponyms contained within a U.S. county geographic boundary. Ads containing multiple toponyms, crossing multiple counties, are assigned fractional ad weights based on the number of county bins a location tag resolves to, given below by the function $geobins$. The $weight$ of any given set of ads A is calculated as:

$$w(A) = weight(A) = \sum_{ad \in A} \frac{1}{geobins(ad)} \tag{1}$$

Formally, prevalence is calculated as follows: given M, the set of all MSM ads for a given location and time interval, the prevalence of a terminology cluster T, at location L, at time window t_i is:

$$prev(T, L, t_i) = \sum_{tag \in L} \frac{w(\{ad \in M_{tag,t_i} : text(ad) \cap T \neq \emptyset\})}{w(\{ad \in M_{tag,t_i}\})} \tag{2}$$