Hierarchical Bayesian Models for Latent Attribute Detection in Social Media

Delip Rao[a] **Michael Paul**[a] **Clay Fink**[b] **David Yarowsky**[a] **Timothy Oates**[c,a] **Glen Coppersmith**[a]
[a]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD 21218
[b]Applied Physics Laboratory, Johns Hopkins University, Laurel, MD 20723
[c]University of Maryland Baltimore County, Baltimore, MD 21250
{delip,mpaul}@cs.jhu.edu, finkcr1@jhuapl.edu, yarowsky@cs.jhu.edu, oates@cs.umbc.edu, coppersmith@jhu.edu

Abstract

We present several novel minimally-supervised models for detecting latent attributes of social media users, with a focus on ethnicity and gender. Previous work on ethnicity detection has used coarse-grained widely separated classes of ethnicity and assumed the existence of large amounts of training data such as the US census, simplifying the problem. Instead, we examine content generated by users in addition to name morphophonemics to detect ethnicity and gender. Further, we address this problem in a challenging setting where the ethnicity classes are more fine grained – ethnicity classes in Nigeria – and with very limited training data.

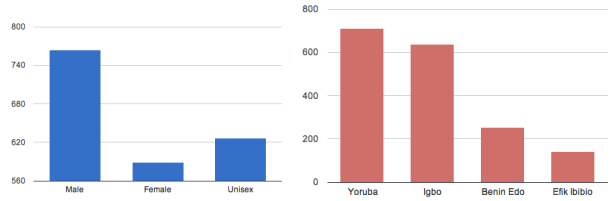
Introduction

The rising popularity of social media websites like MySpace, Facebook and Twitter presents both new opportunities for acquiring intelligence from user-generated content and several challenges. Unlike the traditional web, the content of the social web can actually be tied to the users who generate it. This has important consequences in targeted advertising and personalization. One might like to learn attributes about users like gender, ethnicity, opinions, and other properties and preferences that might be missing in the metadata or that the user chose not disclose. While some outlets like MySpace and Facebook provide opportunities for the user to disclose attributes, influential social networks like Twitter do not even have options to list structured attributes other than an unstructured "Bio" field. We consider learning attributes like gender and ethnicity of users from the text generated by the users and optionally using the self-identified name of the users. Our choice of gender and ethnicity as attributes to learn is inspired by the direct applications of these attributes and the challenges offered by them. While self-identified names are useful in many cases, in some social networks, most notably Twitter, they can be arbitrary and unreliable. Hence it is desirable to have methods that learn from text alone.

We present three different minimally supervised hierarchical Bayesian approaches and compare them using names and content in isolation and in conjunction. Our models perform well under low resource conditions since they directly generate features over names and content. We also show that our models provide a natural way of combining content and names, thereby presenting multi-view learning of latent attributes.

Data

We automatically constructed a dictionary of Nigerian names and their associated gender and ethnicity by crawling baby name sites and other Nigerian diaspora websites (e.g. onlinenigeria.com) to compile a name dictionary of 1980 names with their gender and ethnicity. Although Nigeria has more than 250 ethnic groups, including all tribes and sub-tribes[1], we retain the top 4 ethnicities: Yoruba, Igbo, Efik Ibibio, and Benin Edo. Although the Hausa-Fulani is a populous community from the north of Nigeria, we did not include it as our dictionary had very few Hausa-Fulani names. Further, Hausa-Fulani names are predominantly Arabic or Arabic derivatives and stand out from the rest of the ethnic groups, making their detection easier. For gender we only focused on names that are exclusively male or female as unisex names could be classified either way.



(a) Gender (b) Ethnicity

Figure 1: Distribution of names in crawled dictionaries

Using the Facebook Graph API [2], we collected comments from eleven different English language pages dedicated to Nigerian political figures or causes. A total of 113,527 comments were collected from 94,388 unique users from July, 2010 through the middle of January, 2011. For 78,871 of these users, the gender given on their profile was public.

[1]https://www.cia.gov/library/publications/the-world-factbook/geos/ni.html

[2]http://developers.facebook.com/docs/api

From these, we selected 500 (89 female, 411 male) users at random and, for those whose walls were public, crawled all available wall posts. We did the same for all users that commented on the first set of users' walls. This resulted in a comment graph with 119,523 wall posts or comments from 23,858 users, inclusive of the original set of users. We then crawled the profiles of the new users whose profiles were public, resulting in a total of 23,811 users. Of these 6,364 identified as female, 12,806 as male, and 4,641 had made their gender private. The ethnicity labels for these users were imputed using a linear kernel based SVM classifier[3] that was trained on the dictionary we crawled earlier.

Feature extraction

Since we began with a sparse set of names, we featurized the names to include character n-gram features, up to order 5, from the names. We distinguish between n-grams that appear at the beginning of the name, the end of the name, or in the middle. In addition to names, we also consider text generated by users, like wall posts and comments from Facebook. The text features include standard word unigram and bigrams and a set of socio-linguistic features that was found useful by Rao and Yarowsky (2010) in the context of Twitter for gender identification. A few examples of these features are shown in Table 1. Some statistics about the resulting features in our dataset are shown in Table 2.

FEATURE	*Description/Example*
SMILEYS	A list of emoticons compiled from the Wikipedia
OMG	Abbreviation for 'Oh My God'
ELLIPSES	'....'
REPATED ALPHABETS	E.g. niceeeeee, noooo waaaay
EXCITEMENT	A string of exclamation symbols (!!!!!)
PUZZLED PUNCT	A combination of any number of ? and ! (!?!!??!)
DISFLUENCIES	E.g. Ugh, mmmm, hmmm, ahh, grrr
AGREEMENT	E.g. yea, yeah, ohya

Table 1: A partial list of Socio-Linguistic features

	# feature types	# feature tokens
Name features	54,751	288,378
Content ngram features	690,036	2,731,721
Sociolinguistic features	3273	223,820

Table 2: Feature statistics for the Nigeria dataset

Models

We now present hierarchical Bayesian models that can perform attribute detection under various resource availability conditions:

1. When only a weak prior over names is available via a name dictionary
2. When only content (text) generated by the user is available
3. When both name and content are available

[3]93% 10-fold cross-validation accuracy on dictionary data. We experimented with other kernels; linear kernels worked best.

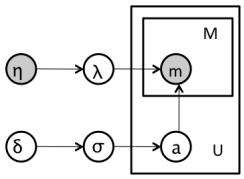


Figure 2: Generative model for latent attribute detection from name-derived features m.

Name Models

The idea behind this model is that each user has an attribute a along with a set of observed name features m. Under this model, user data is generated by first sampling an attribute value a for the user based on some distribution in a population. The value of a can be either observed or hidden: we can keep this value fixed for any users whose attribute value is labeled; we can impute the value of a for users where this is unknown.

The name features are then generated based on the selected value of a. Chang et al (2010) propose a similar model where each a indexes a distribution over first names and a distribution over last names – the distribution over first names must be learned, while the distribution over last names is observed from U.S. Census data. We differ in two ways from this model. First, to overcome the potential issue of name sparsity, we generate an entire set of name features rather than just first and last names. Second, while we do have some prior knowledge about associations between names and attributes, we do not have nearly enough data to use this as a fixed, observed distribution. We will instead encode our name-attribute information in the form of a Dirichlet prior over the distributions, but the model is still free to learn a new distribution that fits the unseen data. The generative story for a corpus of users is as follows:

1. Draw a distribution over attributes $\sigma \sim \text{Dir}(\delta)$.
2. Draw a distribution over name features $\lambda_a \sim \text{Dir}(\eta_a)$ for each attribute a.
3. For each user:
   (a) Sample an attribute $a \sim \text{Mult}(\sigma)$.
   (b) For each name feature i in the user's name, sample a feature value $m_i \sim \text{Mult}(\lambda_a)$.

Figure 2 provides a summary representation of the model with the usual notation of shaded circles representing observed variables and the remaining being latent.

Inference of the hidden variables is straightforward using Gibbs sampling. We can derive a collapsed Gibbs sampler by marginalizing the multinomials out of the equation, requiring us to sample only the variable a for each user u (Griffiths and Steyvers 2004). The sampling equation is:

$$p(a_u|\mathbf{a_{-u}}, \mathbf{m}) \quad (1)$$

$$\propto \quad p(a_u|\mathbf{a_{-u}}, \delta) \prod_{i=1}^{M_u} p(m_{u,i}|\mathbf{a}, \mathbf{m}_{-(u,i)}, \eta_a)$$

$$= \quad \frac{n_a^* + \delta}{n_*^* + A\delta} \prod_{i=1}^{M_u} \frac{n_a^m + \eta_{a,m}}{n_a^* + \sum_k^M \eta_{a,k}}$$

In the sampling equation, the Dirichlet prior effectively acts as a vector of "pseudocounts" which bias the distribution. We set δ to a constant 10.0, while the η values are estimated from our name dictionary. Specifically, the value of $\eta_{a,m}$ is the probability $p(m|a)$ of the feature m appearing in the dictionary for the attribute a, scaled by a factor to increase the size of the effective pseudocounts. We manually set this scaling factor to a large value of cM where M is the size of the name feature set and c is a manually set constant.

Content Models

We construct a similar generative model where the user's attribute value predicts the content features that are generated rather than the user's name features. We add an additional layer of structure to this model such that the word/content features are generated according to a topic model. Topic models such as latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) associate each word token with a latent *topic* variable. Each document is a multinomial mixture of topics, and each topic indexes a multinomial distribution over words. Inferring the hidden topic variables often results in learning word distributions such that topics form semantically coherent word clusters. Topics can be useful features for prediction because they capture collocations and long-range contexts. Additionally, they are useful for qualitative analysis: we can examine the inferred word clusters to see sets of features that may have interesting properties (e.g. some clusters are associated with males versus females).

Specifically, we model a user's topic distribution to have a Dirichlet prior which depends on the value of a. Additionally, we modify the basic LDA model to include a "background" distribution for common words/features that appear independently of the user's topics (Chemudugunta, Smyth, and Steyvers 2006) – this is done to filter out common features that have no predictive value. We will stick with the conventional notation that the topic generates a word/feature w, however note that "words" are actually arbitrary features like content ngrams and sociolinguistic features in our case.

The generative story under this model is:

1. Draw a distribution over attributes $\sigma \sim \text{Dir}(\delta)$.
2. Draw a distribution over words/features $\phi_z \sim \text{Dir}(\beta)$ for each topic z.
3. For each user:
   (a) Sample an attribute $a \sim \text{Mult}(\sigma)$.
   (b) Draw a topic distribution $\theta \sim \text{Dir}(\alpha_a)$.
   (c) Draw a switching distribution $\pi \sim \text{Beta}(\gamma_0, \gamma_1)$.
   (d) For each word/feature i in the user's content messages:
      i. Sample $x_i \in \{0, 1\} \sim \text{Bin}(\pi)$.
      ii. If $x_i = 0$ (the word comes from the background):
          Sample a word $w_i \sim \text{Mult}(\phi_B)$.
      Else if $x_i = 1$ (the word comes from a topic):
          Sample a topic $z_i \sim \text{Mult}(\theta)$.
          Sample a word/feature $w_i \sim \text{Mult}(\phi_z)$.

A collapsed Gibbs sampler can be used for this model as well. In addition to the values for the user attributes a, we must sample values for the variables z and x for each token.

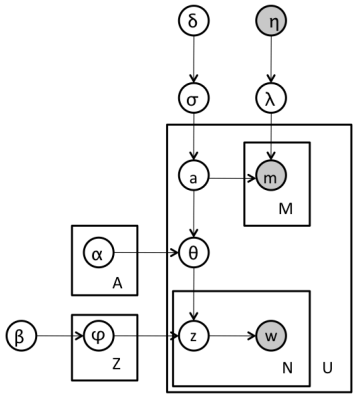


Figure 3: Generative model for combining name features and content features.

For each user, we first sample a conditioned on the current values of z and x, then we resample z and x conditioned on the newly sampled value for a. The update equation is similar to the name model:

$$p(a_u|\mathbf{a_{-u}}, \mathbf{z}, \mathbf{x}) \quad (2)$$
$$\propto \quad p(a_u|\mathbf{a_{-u}}, \delta) \prod_{i:x_i=1} p(z_{u,i}|\mathbf{z}_{-(u,i)}, \alpha_a)$$
$$= \quad \frac{n^*_a + \delta}{n^*_* + A\delta} \prod_{i:x_i=1} \frac{n^z_u + \alpha_{a,z}}{n^*_u + \sum_k^Z \alpha_{a,k}}$$

Sampling the values for z and x are done using standard LDA sampling methods (Griffiths and Steyvers 2004). We also need to learn the α values to make associations between attribute values and topics. Following Pachinko allocation, we update α after each sampling iteration using the update equation given by Li and McCallum (2006).

Combined Name and Content Models

Finally, we consider a simple combination of these two models. The idea is similar: a user's attribute, name features, and content features are jointly generated by first sampling a value for the attribute, then both the name features and content features are independently generated, given a. This is illustrated in Figure 3.

Gibbs sampling for this model is nearly the same; the difference is that now the distribution over the possible values of a to sample from is conditioned on both the name features m and the topic variables z and x.

Experiments & Results

We evaluated our three model variants under different conditions and present some of the results here. Further, we evaluate the name model and name+content model under two settings: 1) When only a weak prior from the dictionary data is present; we call this 'unsupervised' since we don't use any true or imputed labels from Facebook and 2) when both the dictionary prior and some labels from the Facebook data is present; we call this 'semi-supervised' because our models make use of all of the data even though only some of the examples will be labeled. The results in the semi-supervised

setting are reported using 10-fold cross-validation. In addition to the hierarchical Bayesian models, we show results from a smoothed-Naïve Bayes and linear kernel support vector machine implementations (with tuned parameters using a grid search). Two naive baselines are shown in Table 3. They are 1) where labels are assigned uniformly at random (UNIFORM) and 2) where labels are assigned according the empirical prior distribution (PRIOR).

	Gender	Ethnicity
UNIFORM	50.0	25.0
PRIOR	66.8	42.6

Table 3: Naive Baselines for Gender and Ethnicity

We first consider the case where only the dictionary names data is available. The 'unsupervised' for Naïve Bayes and SVM means they were trained with the dictionary features and labels but tested on Facebook users while 'semi-supervised' means that in addition to all dictionary names and labels we also add some of the labels of the Facebook users and test on the remaining users. All results reported are average 10-fold cross-validation accuracies. The results of the name model are shown in Table 4.

Names Only	Unsupervised		Semi-supervised	
	Gender	Ethnicity	Gender	Ethnicity
Naïve Bayes	72.4	**75.1**	75.7	77.2
SVM	70.1	73.3	73.4	76.9
UNHB	**75.8**	63.6	**79.8**	**80.9**

Table 4: Name model results: UNHB is the proposed hierarchical Bayesian name model.

For our 'content only' experiments we report the average 10-fold accuracies in Table 5. CHB, the hierarchical Bayesian model that is an adaptation of the LDA topic model, performs well on gender as opposed to ethnicity when compared to the other classifiers indicating that topics in the social discourse (wall posts and comments) were less indicative of ethnicity than they were of gender.

Content only	Gender	Ethnicity
Naïve Bayes	54.6	**48.7**
SVM	53.9	44.3
CHB	**58.6**	38.4

Table 5: Content model results: CHB is an LDA variant.

Our final model tries to combine names and content by simultaneously generating the name and content features as discussed earlier. The results for this model are shown in Table 6. The unsupervised model experiments (i.e., without using any labels with content features) are only reported for the hierarchical Bayesian model as they are not applicable to SVM and Naïve Bayes. The results in this table show the real benefits of the hierarchical Bayesian approach to feature combination. In the SVM and Naïve Bayes cases, the name features get washed out by high dimensional content features. For SVMs we tried varying the regularization parameter (C) but it did not improve this.

Names + Content	Unsupervised		Semi-supervised	
	Gender	Ethnicity	Gender	Ethnicity
Naïve Bayes	NA	NA	54.0	47.6
SVM	NA	NA	53.9	44.2
NCHB	76.1	63.1	**80.1**	**81.1**

Table 6: Combined model results: NCHB is the proposed hierarchical Bayesian model to combine name and content.

How hard is Nigerian name gender detection?

In order to answer this question we posted around 100 randomly-picked non-unisex first names to 6 annotators who are native Nigerians via Amazon Mechanical Turk and obtained their judgements. We further instructed our annotators to not consult any additional resources like Google and only rely on their domain expertise. The judgements were then compared with the self-identified gender labels from Facebook and we found an average agreement of 0.74 (std = 0.14). This is very close to the proposed hierarchical Bayesian model, our best performing name-only model in an unsupervised setting, with an accuracy of 0.758.

Conclusions

We proposed three different minimally supervised hierarchical Bayesian approaches to latent attribute detection using names and user-generated textual content, in isolation and in conjunction. By utilizing just a small dictionary of names, we showed that our models perform well under low resource conditions because they directly generate features over names and content. Unlike traditional classification approaches in this domain, our models provide a natural way of combining information from user content and name morphophonemics. This can be easily extended to include geotagged location, non-textual content, and other features.

Acknowledgements

This work was partially supported by ONR grant N00014-10-1-0523, and by an NSF Graduate Research Fellowship to the second author.

References

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)* 3:993–1022.

Chang, J.; Rosenn, I.; Backstrom, L.; and Marlow, C. 2010. epluribus: Ethnicity on social networks. In *Proceedings of the International Conference in Weblogs and Social Media (ICWSM)*.

Chemudugunta, C.; Smyth, P.; and Steyvers, M. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *Proceedings of NIPS*, 241–248.

Griffiths, T., and Steyvers, M. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*.

Li, W., and McCallum, A. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the International Conference on Machine Learning*.

Rao, D., and Yarowsky, D. 2010. Detecting latent user properties in social media. In *Proceedings of the NIPS workshop on Machine Learning for Social Networks (MLSC)*.