Discovering Demographic Language Variation

Brendan O'Connor **Jacob Eisenstein** **Eric P. Xing** **Noah A. Smith**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15215, USA
{brenocon,jacobeis,epxing,nasmith}@cs.cmu.edu

Abstract

We propose a Bayesian generative model of how demographic social factors influence lexical choice. We apply the method to a corpus of geo-tagged Twitter messages originating from mobile phones, cross-referenced against U.S. Census demographic data. Our method discovers communities jointly defined by linguistic and demographic properties.

1 Introduction

Even within a single language community, speakers from different backgrounds demonstrate substantial linguistic variation. Salient speaker characteristics include geography [10, 6], race [13], and socioeconomic status [8, 4]; they impact language at the phonological, lexical, and morphosyntactic levels [16]. Sociolinguistics and dialectology feature a strong quantitative tradition of studying the relationship between language and social and geographical identity (e.g., [9, 14]). In general, these approaches begin by identifying both the communities of interest and the relevant linguistic dimensions of variability; for example, a researcher might identify the term "yinz" as characteristic of Pittsburgh dialect [3], and then model its relationship to the socioeconomic status of the speaker. While this approach has a quantitative foundation in modeling the relationship between linguistic and extra-linguistic data, it requires extensive fieldwork and linguistic expertise to identify the "inputs" that are to be analyzed.

In this paper, we propose a new exploratory methodology for *discovering* demographic and geographic language variation from text and metadata. We unite these information sources in a Bayesian generative model, which explains both linguistic variation and demographic features through a set of generative distributions, each of which is associated with a (latent) community of speakers. Thus, our model is capable of discovering both the relevant sociolinguistic communities, as well as the key dimensions of lexical variation.

2 Data

The increasing pervasiveness of social media with structured metadata enables new computational approaches that combine text and demographics. We are particularly interested in the use of GPS-enabled mobile devices to post social media, which connects social computing to the real world. Our dataset is gathered from the microblog website Twitter, via its official API, and consists of an archive of microblog messages which are tagged with the GPS location of the author (we simply use the first recorded GPS location for each author). Authors were selected from a pool of individuals who posted frequently during the first week of March 2010. We use a randomly-selected subset of 4875 authors in this research.

Demographic Variable	Mean	SD	Words with Highest Average
Percent White	52.1%	29.0	leno, fantastic, holy, military, review
Percent African-American	32.2%	29.1	lml, momz, midterms, bmore, fuccin
Percent Hispanic	15.7%	18.3	cuando, estoy, pero, eso, gracias
Percent English speakers	73.7%	18.4	#lowkey, porter, #ilovefamu, nc, atlanta
Percent population in urban areas	95.1%	14.3	odeee, thatt, m2, maddd, mangoville
Percent family households	64.1%	14.4	mangoville, lightskin, iin, af, aha
Median annual income†	$39,045	(26k, 59k)	mangoville, tuck, itunes, jim, dose

Figure 1: List of demographic variables used, selected from 2000 U.S. Census data, along with their mean and standard deviation among authors in the data, and the words with the highest sample-adjusted average values. The procedure for selecting words is described in Section 2; some analysis appears in Section 4. † For the model, income is in log-dollars, but presented above in raw dollars. The SD column shows $(\hat{\mu} \pm \hat{\sigma})$ values computed on a log scale, rescaled back into raw dollars.

Informal text from mobile phones is challenging to tokenize; we adapt a publicly available tokenizer[1] originally developed for Twitter [11], which preserves emoticons and blocks of punctuation and other symbols as tokens. For each user's Twitter feed, we combine all messages into a single "document." The vocabulary is pruned to a size of 5418 by eliminating low-frequency terms. No stopword removal is performed, as the use of standard or non-standard orthography for stopwords may convey important information about the author.

The dataset applied here is adapted from earlier work by Eisenstein et al [5]; for this research, we have extended the corpus with detailed demographic metadata. While it is difficult to identify the demographic attributes of individual speakers, we can cross-reference speaker locations against U.S. Census data (year 2000) to extract aggregate demographic statistics of each user's geographic location.[2] We use the Zip Code Tabulation Areas level of granularity, which partitions the U.S. into 33,178 polygons. Using a standard geospatial tool (PostGIS), we match each author's location to the area that contains it, and use the area's demographics as that author's demographic metadata. The set of features that we consider are shown in Table 1. To give a rough view of word association, for each variable we show the top five words ranked by the sample-corrected average demographic value among authors who use them at least once.[3]

Aggregate demographic statistics must be interpreted with care. If, for example, we find that an author is located in a zip code with 80% Hispanic/Latino population,[4] we do *not* know the likelihood that the author is Hispanic, because the set of Twitter users is not a representative sample of the overall population. Thus, we can only interpret an author's metadata vector as a reflection of the author's surrounding community.

3 Model

We construct a latent variable model that combines the demographic metadata with microblog text from Twitter. The goal is to extract a set of latent sociolinguistic communities which are coherent with respect to both data sources. Our model combines a multinomial distribution over text with a multivariate Gaussian over the demographic statistics; these generative components are unified in a Dirichlet process mixture, in which each speaker has a latent "community" index.

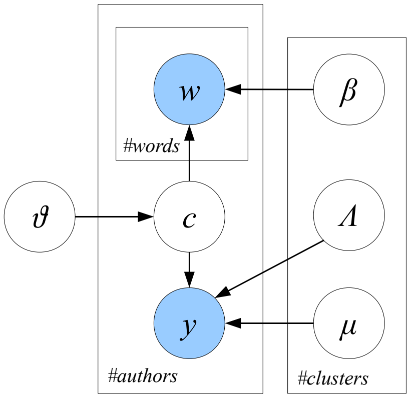
More formally, we hypothesize a generative stochastic process that produces the text and the demographic data for each author. This generative process includes a set of latent variables; we will recover a variational distribution over these latent variables using naïve mean field. The plate diagram for this model is shown in Figure 2. The key latent variable is the community membership of each author, which we write c_d; this variable selects a distribution over both the metadata and

[1] http://tweetmotif.com

[2] Summary Files 1 and 3: http://www.census.gov/support/cen2000.html

[3] We rank by the lower bound of the 95% confidence interval for the mean: $\hat{\mu} - 1.96\,\hat{\sigma}/\sqrt{n}$. Using the raw average always places rare words in the top ranks, which is often due to statistical noise.

[4] In the United States Census, the official ethnonym is *Hispanic or Latino*; for brevity we will use *Hispanic* in the rest of this paper.



c_d	the community membership for author d
y_d	metadata vector for author d
$w_n^{(d)}$	the observed word type for token n of document d
$\boldsymbol{\mu}_i$	mean metadata vector for community i
Λ_i	precision matrix over metadata for community i
$\boldsymbol{\beta}_i$	word distribution for community i
$\boldsymbol{\alpha}$	global prior over topic proportions
$\boldsymbol{\vartheta}$	global prior over communities

Figure 2: Plate diagram for our model of text and demographics, with a table of all random variables. The document indices in the figure are implicit, as are the priors on $\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\beta}$.

the text. Each distribution over metadata is a multivariate Gaussian with parameters $\boldsymbol{\mu}$ and Λ; each distribution over text is a multinomial with parameter $\boldsymbol{\beta}$. Overall, we can describe the generative process as follows:

- Draw the community proportions $\boldsymbol{\vartheta}$ from a stick-breaking prior,
- **Generate the community distributions.** For each community i,
  - Draw the metadata mean $\boldsymbol{\mu}_i$ from a multivariate Normal prior,
  - Draw the metadata precision matrix Λ_i from a Wishart prior,
  - Draw the word distribution $\boldsymbol{\beta}_i$ from a Dirichlet prior,
- **Generate the text and metadata.** For each author d,
  - Draw the community c_d from the distribution $\boldsymbol{\vartheta}$,
  - Draw the metadata $\boldsymbol{y}_d$ from a Gaussian with mean $\boldsymbol{\mu}_{c_d}$ and precision Λ_{c_d},
  - Draw the bag of words $\boldsymbol{w}_d$ from the multinomial $\boldsymbol{\beta}_{c_d}$,

We apply a naïve mean field to recover a variational distribution Q over the random variables in this model [1]. Specifically, we place variational distributions over all latent variables of interest, and iterate between updates of the community parameters ($\boldsymbol{\mu}, \Lambda, \boldsymbol{\beta}, \boldsymbol{\vartheta}$) and updates for the authors' community memberships ($\boldsymbol{c}$). A complete description of mean field inference for multivariate Gaussian mixtures is presented by Penny (2001). Our inference differs in that (a) each mixture component is also responsible for generating the text, and (b) we place a non-parametric stick-breaking prior over the community proportions $\boldsymbol{\vartheta}$, which automatically selects the appropriate number of communities, using the truncated stick-breaking approximation [2]. For efficiency, we initialized by running the Dirichlet process mixture model on the demographic data alone; we selected the number of active clusters from this initialization, and used this as the truncation level on the text+demographics data.

4 Analysis

Figure 1 shows the words associated with large values for each demographic feature, ranked by sample-corrected averages (described above). This is informative and should be viewed as an exploratory method in its own right; however, correlations between demographic variables make it difficult to disentangle the underlying relationships between demography and lexical frequencies.

In contrast, Figure 3 summarizes all of the sociolinguistic clusters identified by our model. All clusters are shown. Each row shows a cluster that corresponds to a distribution over demographic information, along with a set of characteristic terms as chosen by likelihood ratio. This cluster analysis allows us to associate each term with a complete demographic profile.

Several of the top terms refer to subjects which attracted only an ephemeral interest, and would likely not appear in a dataset taken from a longer timespan. The term *19th* refers to an event on

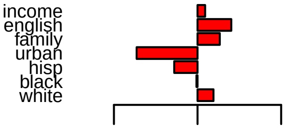
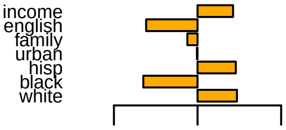
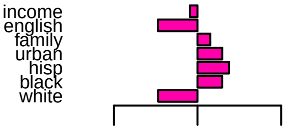
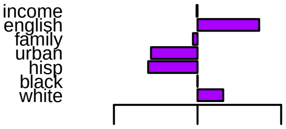
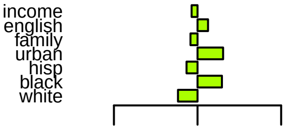
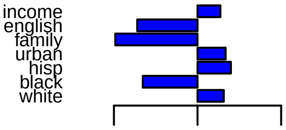
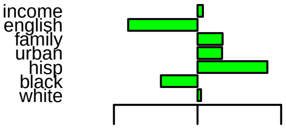
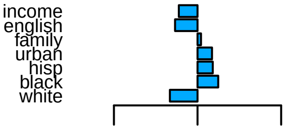
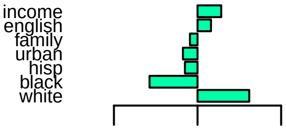
	Mean Vector	Top Words
Cluster 1		rsvp, ent, guest, blvd, broadcasting, details, bash, lls, retweet, ——, #free, —-, hosting, pow, vibe, 31, vol, —–, feat, 2nite
Cluster 2		en, de, el, que, es, por, se, un, los, pero, una, como, para, lo, del, te, si, eso, la, tu
Cluster 3		ii, dha, yu, uu, yuu, dhat, lols, lolss, lml, qo, qot, w—, myy, iim, qet, yuh, smhh, niqqa, buh, &&
Cluster 4		#ilovefamu, lmbo, grind, official, awards, #lowkey, #famusextape, jake, track, spirit, #thatisall, mental, famu, praying, studying, you're, bible, midterm, joy, awesome
Cluster 5		lls, jawn, neighbors, joints, nivea, #famusextape, sextape, #epicfail, cuddle, broad, midterms, jeezy, #thatisall, basic, nigga, waited, tmobile, menu, bcuz, famu
Cluster 6		gimmie, hosted, —, download, b-day, dl, limit, drake, mix, dj, mc, salute, //, #unotfromthehoodif, ft, exclusive, birthday, models, -, lab
Cluster 7		:], ;], -_-, ^, bahaha, :d, papi, ^_^, ily, aha, =], fck, hahah, lovee, ew, yess, :/, #urparentsever, mangoville, jamaica
Cluster 8		dats, dat, dis, wat, da, watz, dey, wats, den, gud, wen, gravity, niggaz, jus, der, fuk, rite, dem, tha, dese
Cluster 9		rare, 19th, olive, simply, adam, agent, coffee, obama, awesome, 400, hockey, leno, thomas, worked, pentagon, #fb, tone, presents, larry, peppers

Figure 3: Demographic mean vectors and most salient words per cluster. Demographic variables are shown on a normalized scale; zero indicates the population mean, and the axis tick marks denote ± 1 standard deviation; see Table 1 for their values. For each cluster, words shown are the top-20 most highly ranked by the ratio of topic probability against background probability: $\frac{\beta_k[w]}{\tilde{p}(w)}$.

March 19, 2010 that was a frequent subject of conversation in this dataset. The term *olive* usually refers to the Olive Garden restaurant; *mangoville* refers to a restaurant in New York City. Florida A&M, a historically-black university, appears in the terms *#ilovefamu* and *#famusextape*. Terms that start with hashtags (e.g., *#epicfail*, *#thatisall*) often represent "trends" that are shown on all users' Twitter pages; many users participate by adding their own commentary on such tags [7]. The topic lists also contain several names, including *leno*, *obama*, and the musicians *drake* and *jeezy*.

One cluster contains exclusively Spanish words. These words exhibit an extremely strong mutual association; authors that use even a few Spanish words are unlikely to use any words in English. This cluster is relatively diffuse with regard to demographic data, and would not be detected without recourse to the linguistic properties of the speakers. Note that while this cluster contains a high proportion of Hispanics, it also appears to contain an above-average number of white speakers. We see two potential explanations: the speakers in this cluster may come from mixed white-Hispanic neighborhoods, or the individual authors may identify as both ethnicities.

We see a number of characteristic phenomena which are characteristic of computer-mediated communication, including emoticons, phonetic spelling, and abbreviations [15]. Emoticons (e.g., *:]*) are grouped in cluster 7, which contains many Hispanics and is above-average with respect to income. Phonetic spelling is used in clusters 3 and 8, which are the two lowest-income clusters and contain the fewest whites. The other group with an above-average number of blacks is cluster 5, and the associated language is somewhat more standardized. Relative to clusters 3 and 8, cluster 5 is wealthier, more urban, and contains more whites and fewer Hispanics.

Clusters 1, 4, and 9 are substantially less urban than the other clusters, and they contain the most standard English words. Of particular interest is cluster 4, which is the only non-urban cluster with lower-than-average income; the fact that this cluster is still relatively standard may provide hint to the relative importance of urbanity and wealth with respect to relative frequency of standard and "vernacular" language.

5 Conclusion

The relationship between language and social identity has traditionally been studied with respect to micro-level phenomena, such as phonological features or individual words. In this paper, we take a more holistic approach, using a generative model that operates on authors' entire microblogging history. Our model extracts sociolinguistic communities that are coherent with respect to both demographic metadata and text; moreover, it identifies individual terms that are especially characteristic of these communities in social media. In the future, we plan to test the ability of our model to predict authors' demographic attributes from raw text.

Acknowledgments

We would like to thank Qirong Ho for invaluable assistance. This research was enabled by Google's support of the Worldly Knowledge project at CMU, AFOSR FA9550010247, ONR N0001140910758, NSF CAREER DBI-0546594, NSF IIS-0713379, and an Alfred P. Sloan Fellowship.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2006.

[3] M. R. Dressman. Redd up. *American Speech*, 54(2):141–145, 1979.

[4] P. Eckert. *Jocks and Burnouts: Social Categories and Identity in the High School*. Teachers College Press, 1989.

[5] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model of geographic lexical variation. In *Proceedings of EMNLP*, 2010.

[6] H. Kurath. *A Word Geography of the Eastern United States*. University of Michigan Press, 1949.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of WWW*, 2010.

[8] W. Labov. *The Social Stratification of English in New York City*. Center for Applied Linguistics, 1966.

[9] W. Labov, editor. *Locating Language in Time and Space*. Academic Press, 1980.

[10] W. Labov, S. Ash, and C. Boberg. *The Atlas of North American English: Phonetics, Phonology, and Sound Change*. Walter de Gruyter, 2006.

[11] B. O'Connor, M. Krieger, and D. Ahn. TweetMotif: Exploratory search and topic summarization for twitter. In *Proceedings of ICWSM*, 2010.

[12] W. D. Penny. Variational Bayes for d-dimensional Gaussian mixture models. Technical report, University College London, 2001.

[13] J. R. Rickford. *African American Vernacular English*. Blackwell, 1999.

[14] S. A. Tagliamonte. *Analysing Sociolinguistic Variation*. Cambridge University Press, 2006.

[15] S. A. Tagliamonte and D. Denis. Linguistic ruin? LOL! Instant messanging and teen language. *American Speech*, 83, 2008.

[16] W. Wolfram and N. Schilling-Estes. *American English: Dialects and Variation*. Basil Blackwell, second edition, 2006.