Detecting Latent User Properties in Social Media

Delip Rao*
Dept. of Computer Science
Johns Hopkins University
delip@cs.jhu.edu

David Yarowsky
Dept. of Computer Science
Johns Hopkins University
yarowsky@cs.jhu.edu

Abstract

The ability to identify user attributes such as gender, age, regional origin, and political orientation solely from user language in social media such as Twitter or similar highly informal content has important applications in advertising, personalization, and recommendation. This paper includes a novel investigation of stacked-SVM-based classification algorithms over a rich set of original features, applied to classifying these four user attributes. We propose new sociolinguistics-based features for classifying user attributes in Twitter-style informal written genres, as distinct from the other primarily spoken genres previously studied in the user-property classification literature. Our models, singly and in ensemble, significantly outperform baseline models in all cases.

1 Introduction

With the emergence of social media portals there is an abundance of user generated content on the Web. Furthermore, the popularity of websites like Twitter, MySpace and Facebook is growing unabated. Twitter alone has over 10 million global users (8 million in the US) when compared to around 6 million for Facebook[1]. This has important consequences in targeted advertising and personalization. However, unlike Facebook or MySpace, Twitter has limited metadata available about its users. Important attributes of the user such as age and gender that are directly useful for providing personalized services are not typically available in profiles or available metadata. Further, one might be interested in knowing other user attributes, such as ethnicity, opinions and other properties and preferences that the user might not disclose.

While there has been previous work on discovering latent speaker/author attributes from conversational genres, including telephone speech (e.g. Garera and Yarowsky [7]) and audio recordings (e.g. Bocklet et. al [1]), similar work on author-property discovery in microblogs such as Twitter is lacking in the literature, and the microblog genre has sufficient unique interesting properties to make investigation of this genre worthwhile. The goal of this paper is to automatically discover some of these *latent* (i.e. not-overtly-stated) attributes of the users using status messages (or 'tweets'), the social network structure, and communication behavior of the users. We also explore additional latent user properties not previously predicted and sociolinguistic features that were not previously explored. Our first finding reveals that the status message content is more valuable in inferring latent author attributes relative to some properties of social network structure or communication behavior. This is in contrast to prior findings in telephone conversational speech where discourse factors have been proven more effective.

Since this a first-of-a-kind application of latent-author-attribute discovery to various author attributes on Twitter, we had to build our own annotated data for the attributes concerned. We then treat this as a supervised classification problem, although semi-supervised variations are definitely possible and are a topic of our ongoing work. From the data, we derive tweet-content based models in two ways –

*Contact author

[1] http://www.allfacebook.com/2009/03/twitter-facebook-comparison/

1) using lexical-feature-based approaches and 2) by extracting and utilizing sociolinguistics-inspired features. Previous work by Garera and Yarowsky [7] showed the importance of sociolinguistic models in classifying latent speaker attributes in conversational speech transcripts. Here we show the effectiveness of these and additional novel features for microblogs. It should be noted that the sociolinguistic features used in [7] are substantially different from the ones utilized here given the significant differences in genre, medium and style, and constraints of the Twitter data relative to conversational speech.

2 Description of Data and Latent Attributes

We investigate and evaluate performance on three biographic author attributes – gender, age, and regional origin – and one personalization attribute – political orientation. This is limited only by our current access to ground-truth data for evaluation, and using the same approach described in this paper one can learn a diverse variety of other attributes including dietary preferences (vegetarian vs. non-vegetarian), sexual orientation (homosexual vs. heterosexual), student status (student vs. non-student) and so on constrained only by available training and evaluation data.

For each of these tasks we built a dataset by focused crawling and manual annotation. For each attribute a 'seed' set of related users is collected and from these seeds other potential candidates are explored in a breadth-first manner via the Twitter follower network[2]. We further ignore all candidates with a high follower count as these tend to celebrities or organizations. Each of the remaining candidates were manually annotated by two independent annotators. The annotators were not exposed to the experimental methods used or the modeling approaches adopted. To avoid a label bias problem we constrained the number of users in each class to be similar. We describe the crawling approaches and the sources used for each of the attributes in detail. All of the manually-annotated Twitter data sets developed in this work will be made available as a new shared resource to the research community.

Gender: The task here is to detect if a Twitter user is male or female simply by the content and behavior of their postings. While many Twitter users use their real name, which can reveal their gender, many choose names such as *graywolf* or *jellyfish* that do not convey gender. To make this work maximally general to all Twitter users, and to informal communication without associated gender-marked user handles, we ignore user names in this classification. Instead we are interested in knowing if male and female Twitter users can be distinguished exclusively from the content and style of their writing.

For gender, the seed set for the crawl came from initial sources including sororities, fraternities, and male and female hygiene products. This produced around 500 users in each class.

Age: While age is a real valued attribute, given the general lack of ground truth for user age, in this paper we employed a labor-intensive manual investigation and classification of users into two major demographic pools, users who are below 30 and users who are above 30. This binary categorization is nevertheless useful for user modeling and understanding. Age is also a difficult attribute to learn. Not only does it change constantly, age-sensitive communication behavior differs based on numerous socioeconomic variables, and there is no well known indicator for age on Twitter.

The crawls for age categories were seeded by looking for twitter lists for initial seed topics including "baby boomers", "young moms" or searches for terms such as "junior" or "freshman" in the user description field. This seed set was expanded by adding all the followers and crawling their profiles. The ground-truth labels for evaluation were then derived by manually looking up the MySpace, LinkedIn profiles, blogs, and/or homepages that many of these users had listed along with their Twitter profile. This yielded around 1000 users in each class.

Regional Origin: The regional origin of a user often correlates with a user's dialect, such as whether they are from the South or NorthEast Urban USA. There has also been interest in classifying speakers of English by their national origin. We combine these into the refined task of classifying whether someone writing in English from India is from Southern or Northern India.

In order to highlight English dialectal differences, an initial seed list of of posts from 3 south and 3 north Indian cities was compiled focusing on cities with relatively low rates of migration (unlike Bangalore or Mumbai) to minimize noise in the data, just as one would exclude Atlanta when devel-

[2]We used the Twitter4J Java API.

oping training data for Southern US dialect. A Twitter search on these cities were used to harvest an initial candidate set of users. The candidate set was further expanded by adding the followers of the users. Finally, the profiles of all these candidates were manually annotated as north or south Indian by two native Indian-language-speaking annotators. This yielded around 200 users per category. We also explored the possibility of using the Twitter GeoLocation API with little success in coverage, possibly due to the low proliferation of GPS enabled mobile devices in India and our intentional exclusion of the large metropolitan cities.

Political Orientation: We investigated Twitter users from the US and tried to distinguish between users with Republican/conservative leanings vs. those with typically Liberal/Left/Democratic leanings. Our seed set for this crawl was populated by looking at twitter lists for the National Rifle Association (NRA), groups like "Expose Liberals", keyword searches like "support Palin" or "proud democrat" and hashtags related to current news events – eg. `#ISupportArizona` for the Arizona Anti-Immigration Law and then iterating. Finally each of the derived candidates was carefully examined by looking into the collected tweet messages and manually the user as the prototypical label "Democrat" or "Republican" as shorthand based on the observed positions. This yielded around 500 users per category. Our goal is to discover if a computer can make this classification automatically.

Overall details regarding the data we collected and annotated are shown in Table 2. In each case, we assumed that a user's Gender, Age, Political Orientation and Regional Origin are stable (e.g. have a single value across all the collected tweets from the user).

3 Classification Models

For the purpose of these experiments, we generalized the four user-property classification tasks as binary classification problems, and built separate binary classifiers for each attribute using Support Vector Machines (SVM) [9]. The features for these classifiers depend on the models being used but all features are derived from the tweet messages alone as our goal was to study the efficacy of language content processing in identification of latent user attributes from observed textual communications. Below we describe three classification models: 1) Sociolinguistic-feature models, 2) Ngram-feature models, and 3) a stacked model that combines the results from the previous two models. We also investigated the use of communication features like reply-rate, follower-followee ratio and discovered that these features, although relevant in speech [7], did not help in latent attribute detection[3].

SocioLinguistic-feature Models: Research in sociolinguistics (see Macaulay [11]) has demonstrated the differences in lexical choice and other linguistic features in discourse conditioned on age, gender, and social class. For example, in speech it is well known that certain utterances like "umm", "uh-huh", and back channel responses like laughter and lip smacking are more prevalent among female speakers than their male counterparts. Mean utterance length and conversational dominance have also been productively employed features.

The absence of prosodic cues in Twitter, the fixed utterance size limitations, and nature of tweeting vs. speech conversation demands that the traditional sociolinguistic cues be re-examined for this genre. We found communication on Twitter and other online media provides valuable sociolinguistic cues that help identify latent attributes. However, the nature of the cues are different. For instance, a very peculiar cue is the presence of a sequence of exclamation marks, which empirically is very indicative of a female user in our dataset. Also we found the presence of certain kinds of emoticons (like `<3`) is also a strong indicator of female users. People laugh differently on Twitter as well. While women `LOL`, men tend to `LMFAO`. These cues are not necessarily restricted to gender. For example, the lexical choices can certainly distinguish between young and old users. The use of address terms like 'dude' and 'bro' almost certainly indicate a younger user. Similarly, we found the older users more articulate, i.e. they use meaningful sentences within character limits as opposed to the more inarticulate choice of using ellipses (....) among many younger users. A complete listing of socio-linguistic features extracted is shown in Table 1. These features are extracted from status messages. For all experiments we only considered messages that were either broadcasts or replies and ignored the 'retweet' (RT) messages since the authorship of this text cannot be attributed to the user. Ideally one would derive (or learn) sociolinguistic cues on a per-attribute basis, but for model generality we utilized and evaluated the same set of sociolinguistic features for all user attributes.

[3]More details can be found in our tech-report: URL withheld for blind review.

FEATURE	*Description/Example*	*#female/#male*
SIMLEYS	A list of emoticons compiled from the Wikipedia.	3.5
OMG	Abbreviation for 'Oh My God'	4.0
ELLIPSES	'...'	1.5
REPATED ALPHABETS	E.g. niceeeeee, noooo waaaay	1.4
EXCITEMENT	A string of exclamation symbols (!!!!!)	2.0
PUZZLED PUNCT	A combination of any number of ? and ! (!?!!??!)	1.8
DISFLUENCIES	E.g. Ugh, mmmm, hmmm, ahh, grrr	1.8
AGREEMENT	E.g. yea, yeah, ohya	0.7

Table 1: A partial list of Socio-Linguistic features

The templates from Table 1 and additional features described later resulted in 3774 unique instantiated feature types. The extracted features were used to learn an SVM based binary classifier; we call this model *socling*. All features are uniformly weighted but the presence of a feature multiple times in a user's tweet stream is accounted for in the real-valued count features. We employed the SVMLite package[4]. Examples and discussion of some instantiated features and their model statistics follow in Section 4.

Ngram-feature Model: Boulis and Ostendorf [2] first proposed the use of SVM-based models for identification of gender from telephone conversations. Garera and Yarowsky [7] further showed the utility of these models in a wider range of conversational data. We thus utilized this as one class of approaches for Twitter classification by deriving the unigrams and bigrams of the tweet text. The text is first segmented and normalized to preserve emoticons and other punctuation sequences as first-class lexical items (they are traditionally deleted). Emoticons serve as important features as explained earlier. Digits are normalized and text lowercased. The unigrams and bigrams thus generated are weighted by their normalized term frequency – we found TFIDF did worse for this task on a development set.

Stacked Model: Finally, we employed a stacked model to do simple classifier stacking. We utilized another SVM for this task, but its features are the predictions from the Ngram-feature and SocioLinguistic models along with their prediction weights.

4 Evaluation

We evaluate the different models described above by dividing each dataset into training, development, and test sets. The development set was used to set the parameters of the SVM including the choice of kernel. Our experiments did not find other kernels to perform any better than the linear kernel for all classification tasks here. By design, we kept the number of examples in each class to be same. So an uninformative prior (chance) would yield only 50% accuracy. Table 2 shows the classification results for the latent attributes for all the models described. Due to limited space we only present important results here and refer the reader to detailed experiments, extensive feature analysis, and discussions to our tech-report [5].

	Gender	Age	Orientation	Region
Number of Users Per Class	500	1000	200	500
Number of Tweets	405151	2367631	1986392	497337
NGram Model Feature Dimension	1256558	4908979	4398829	1752420
Model	*Accuracy*			
Baseline Model (Prior)	50.00	50.00	50.00	50.00
Socolinguistic Feature Model	71.76	69.44	**77.08**	63.37
Ngram-based Feature Model	68.70	73.09	72.92	**82.84**
Stacked Model	**72.33**	**74.11**	73.77	80.19

Table 2: Dataset description and classification accuracy for the various attributes.

Gender: The sociolinguistic model performs better than the lexical ngram model from the status text alone. This indicates the effectives of the sociolinguistic features listed in Table 1. Further the stacked model gives a slight improvement over the *socling* model. We found that female users were 3.5 times more likely to use emoticons than their male counter parts. This is in line with the findings in speech research where laughter in conversational speech is more frequently associated with a female speaker. Men and women also differ in the kind of emoticons used. We found similar gender based differences in the use of elipses, character repetition (like 'niceee'), repeated exclamation, puzzled punctuation and other features.

It should be noted that gender differences in the usage of some of these features are also conditioned on the age of the users. We did not provide any explicit instruction to our annotators to balance

[4]http://svmlight.joachims.org/

[5]URL withheld for blind-review

different age groups and our focused crawling setup for gender managed to get a large proportion of younger users.

We observed that the words following "my" have particularly high value as predictive features for biographic attributes because they distinguish properties of the individual (e.g. "my wife", "my dress", "my gf") from those of other individuals (e.g. "your wife", "his wife"). And while all of these variant bigrams are included in the ngram model, we observe empirically that performance is improved when the *socling* model is augmented to include the instantiated template `my` followed by any word to provide special status to these self-referential features. We also noticed a higher ratio of disfluencies (`oh, hmm, ugh ..`) and lower ratio of agreement expressions (`yeah, yea, ...`) between female and male users.

Male		Female		Democrat		Republican	
my_zipper	1	*my_zzz*	1	my_youthful	1	*my_zionist*	1
my_wife	0.96	*my_yogurt*	1	my_yoga	1	*my_yuengling*	1
my_gf	0.96	*my_yoga*	1	my_vegetarianism	1	*my_weapons*	1
my_nigga	0.91	*my_husband*	0.98	my_upscale	1	*my_walmart*	1
my_want	0.83	*my_bf*	0.97	my_tofurkey	1	*my_trucker*	1
my_beer	0.78	*my_prof*	0.95	my_synagogue	1	*my_patroit*	1
my_shorts	0.67	*my_daddy*	0.94	my_lakers	0.93	*my_lsu*	1
my_jeep	0.6	*my_research*	0.93	my_gays	0.8	*my_blackeberry*	1
my_woman	0.5	*my_gosh*	0.92	my_feminist	0.67	*my_redneck*	0.89
my_vegas	0.5	*my_bff*	0.88	my_sushi	0.6	*my_marine*	0.82

Table 3: Top possessive features (my_xxx) for `GENDER` and political `ORIENTATION` attributes

Age: From Table 2, the ngram model performs better than the sociolinguistic model and the stacked model improves over both. The reason for the relatively poor performance of the sociolinguistic features for age compared to the status-ngram-based models is because we did not observe major differences in informal writing styles of older and younger users in Twitter (The age corpus was also balanced for gender, and much older users are typically not on twitter at all.) For instance, the use of emoticons was almost identical. But alphabetic character repetition was 30% more likely in younger twitter users. This suggests further investigation in attribute specific sociolinguistic cues.

Regional Origin: The Regional Origin of a user is the hardest attribute to learn. In fuller generality, it's an open class attribute and often gets complicated due to mixed origins and nurture vs. environmental effects on dialect and usage. Our setup was to identify the regional origin of Indian English-writing authors based on their dialectical differences in English tweets. Our experiments (see Table 2) shows the effectiveness of the sociolinguistic-feature models over ngram-based models.

In order to determine if region-specific name or location usage (e.g. common locations and given/surnames of the region) rather than strictly language usage/style was affecting performance, we further filtered out people and place proper names, using a gazetteer and namelist, from the tweet messages to see to what extent the ngram-feature model relied on those proper-name features (capitalization was insufficient by itself given the frequently all-lower-case text in Twitter). We did not measure any significant changes to the results, which shows that these results are driven primarily by style and common noun/verb lexical usage variations rather than region-specific names. Our general set of sociolinguistic features listed in Table 1 seemed to discriminate well between the north and south Indian users. Our annotators were instructed to balance male and female users in this data.

Political orientation: The last attribute we considered is the personalization or user-preference attribute of political orientation. This attribute is of course not formally intrinsic to the user and could change over time (although it rarely does in short time windows). Nevertheless, learning user preferences has important consequences in recommendation and targeted advertising. Learning political orientation from congressional debate texts has been explored earlier [14]. However doing so on microblogging environments is challenging due to limited availability of text, and unlike congressional debates, user twitter streams are topically heterogenous and discuss political topics very sporadically, if at all. However the presence of other benign indicators might reveal the users true intentions or affiliations. In our corpus we've noticed that even if the user did not discuss politics explicitly the presence of possessive's such as "my_tofurkey" strongly indicated democrat while "my_handgun" almost always indicated republican. The generic sociolinguistic features fared poorly compared to the model derived from status ngrams indicating the need for further research in sociolinguistic cues for political orientation.

5 Related Work

Research is sociolinguistics has explored the effects of gender, age, social class, religion, education and other speaker attributes in conversational discourse and monologue. Earliest work in this area by Fischer [6] and Labov [10] involved studying morphological and phonological features respectively. Of all the attributes, gender has been extensively studied presumably because of its anthropological implications (see Coates [4] and Eckert and McConnell-Ginet [5]).

Early computational models for detecting gender from text were primarily interested in formal text (E.g., Singh [13] and Herring and Paolillo [8]) although standard-prose blogs were considered in later works (Burger and Henderson [3]; Nowson and Oberlander [12]).

Recent state-of-the-art work in this field is by Garera and Yarowsky [7], who extend the ngram based models proposed in Boulis and Ostendorf [2] with sociolinguistic features and show applicability to a variety spoken conversational transcripts and more formal enron email corpus. The automatic classification of diverse user attributes from highly informal microblog postings on the other hand has been essentially unexplored, however, and unlike conversational speech offer several challenges including limited size of the tweets, their informal nature, lack of prosodic cues unlike speech, and departure from traditional sociolinguistic cues. The benefits of detecting latent attributes of users on a growing platform like Twitter with a wide reach is immense in business and government endeavors. Our work is the first multi-attribute report of performance on this new genre.

6 Conclusions

This paper has presented a novel set of features and approaches for automatically classifying latent user attributes including gender, age, regional origin, and political orientation solely from the user language of informal communication such as Twitter. Success in this enterprise has important applications in advertising, personalization, and recommendation. The paper has included a novel investigation of stacked-SVM-based classification algorithms over a rich set of original features, applied to classifying these four user attributes. We proposed new features for classifying user attributes in Twitter-style informal written genres as distinct from the other primarily spoken genres previously studied in the user-property classification literature. Our models, singly and in ensemble, significantly outperform baseline models in all cases.

References

[1] T. Bocklet, A. Maier, and E. Nöth. Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines/regression. In *Proceedings of Text, Speech and Dialogue (TSD)*, pages 253–260, Berlin, Heidelberg, 2008. Springer-Verlag.

[2] C. Boulis and M. Ostendorf. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the ACL*, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[3] J. Burger and J. Henderson. An exploration of observable features related to blogger age. In *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium*, 2006.

[4] J. Coates. *Language and Gender: A Reader*. Blackwell Publishers, 1998.

[5] P. Eckert and S. McConnell-Ginet. *Language and Gender*. Cambridge University Press, 2003.

[6] J. Fischer. Social influences on the choice of a linguistic variant. In *Proceedings of Word*, 1958.

[7] N. Garera and D. Yarowsky. Modeling latent biographic attributes in conversational genres. In *Proceedings of ACL-IJCNLP*, pages 710–718, 2009.

[8] S. Herring and J. Paolillo. Gender and genre variation in weblogs. In *Journal of Sociolinguistics*, 2006.

[9] T. Joachims. *Learning to Classify Text using Support Vector Machines*. Kluwer, 2002.

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

[11] R. K. Macaulay. *Talk that counts: Age, Gender, and Social Class Differences in Discourse*. Oxford University Press, 2005.

[12] S. Nowson and J. Oberlander. The identity of bloggers: Openness and gender in personal weblogs. In *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium*, 2006.

[13] S. Singh. A pilot study on gender differences in conversational speech on lexical richness measures. In *Literary and Linguistic Computing*, 2001.

[14] M. Thomas, B. Pang, and L. Lee. Get out the vote: determining support or opposition from congressional floor-debate transcripts. In *EMNLP '06*, 2006.